EVALUATING TRUST AND EXPLAINABILITY FOR DEEP LEARNING MODELS

QUANTIFYING TRUST IN DEEP LEARNING WITH OBJECTIVE EXPLAINABLE AI METHODS FOR ECG CLASSIFICATION

BY

KASHIF SIDDIQUI, B.Eng.

A THESIS
SUBMITTED TO THE DEPARTMENT OF SCHOOL OF BIOMEDICAL ENGINEERING
AND THE SCHOOL OF GRADUATE STUDIES
OF MCMASTER UNIVERSITY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF APPLIED SCIENCE

Master of Applied Science (2020)
(School of Biomedical Engineering)

McMaster University
Hamilton, Ontario, Canada

TITLE:	Quantifying Trust in Deep Learning With Objective Explainable AI Methods for ECG Classification
AUTHOR:	Kashif Siddiqui B.Eng. (Electrical & Biomedical Engineering), McMaster University, Hamilton, Canada
SUPERVISOR:	Dr. Thomas Doyle
NUMBER OF PAGES:	xxi, **??**

Lay Abstract/ Thesis statement

The goal of this thesis was to develop a framework of how trustworthiness can be improved for a variety of stakeholders in the use of AI in medical applications. Trust was broken down into basic elements (Explainability, Verifiability, Fairness & Robustness) and 'Explainability' was further explored. This was done by determining how explainability (offered by XAI methods) can address the needs (Accuracy, Safety, and Performance) of stakeholders and how those needs can be evaluated. Methods of comparison (similarity, stability, and novelty) were developed that allow an objective evaluation of the explanations from various XAI methods using repeatable metrics (Jaccard, Hamming, Pearson Correlation, and TF-IDF). Combining the results of these measurements into the framework of trust, work towards improving AI trustworthiness and provides a way to evaluate and compare the utility of explanations.

Abstract

Trustworthiness is a roadblock in mass adoption of artificial intelligence (AI) in medicine. This thesis developed a framework to explore the trustworthiness as it applies to AI in medicine with respect to common stakeholders in medical device development. Within this framework the element of explainability of AI models was explored by evaluating explainable AI (XAI) methods. In current literature a litany of XAI methods are available that provide a variety of insights into the learning and function of AI models. XAI methods provide a human readable output for the AI's learning process. These XAI methods tend to be bespoke and provide very subjective outputs with varying degrees of quality. Currently, there are no metrics or methods of objectively evaluating XAI outputs against outputs from different types of XAI methods. This thesis presents a set of constituent elements (similarity, stability and novelty) to explore the concept of explainability and then presents a series of metrics to evaluate those constituent elements. Thus providing a repeatable and testable framework to evaluate XAI methods and their generated explanations. This is accomplished using subject matter expert (SME) annotated ECG signals (time-series signals) represented as images to AI models and XAI methods. A small subset from all available XAI methods, Vanilla Saliency, SmoothGrad, GradCAM and GradCAM++ were used to generate XAI outputs for a VGG-16 based deep learning classification model. The framework provides insights about XAI method generated explanations for the AI and how closely that learning corresponds to SME decision making. It also objectively evaluates how closely explanations generated by any XAI method resemble outputs from other XAI methods. Lastly, the framework provides insights about possible novel learning done by the deep learning model beyond what was identified by the SMEs in their decision making.

This thesis is dedicated to my darling fiancée and my loving family

I could not have done this without you.

Acknowledgements

I would like to acknowledge the mentorship and invaluable guidance of Dr. Thomas Doyle, through every stage of my thesis.

Contents

List of Figures

List of Tables

Abbreviations

Abbreviations

AE	Auto Encoder
AFIB	Atrial Fibrillation
AI	Artificial intelligence
AUC	Area Under the Curve
BERT	Bidirectional Encoder Representations from Transformers
BW	Box-and-Whisker plot
CMA	Canadian Medical Association
CNN	Convolutional Neural Network
DARPA	Defense Advanced Research Projects Agency
DC	Direct Current
DCN	Deep Convolutional Network

DL	Deep Learning
DNN	Deep Neural Network
ECG	Electrocardiogram
E_xai	XAI Trustworthiness Vector
GC	GradCAM
GC++	GradCAM++
GL	Global (vs. see LO)
GSVT	Generalized Supra-ventricular Tachycardia
HC	Healthcare
IMG	Image
IN	Inherent (vs. see PH)
IQR	Interquartile Range
K_n	Knowledge (Novelty)
LO	Local (vs. see GL)
LOESS	Local polynomial regression
MA	Model Agnostic (vs. see MS)
MITBIH	Massachusetts Institute of Technology-Beth Israel
ML	Machine Learning

MS	Model Specific (vs. see MA)
NHS	National Health Service
NLM	Non Local Means
NN	Neural Network
OTSU	Thresholding method named after Nobuyuki Otsu
PCr	Pearson Correlation
PH	Post Hoc (vs. see IN)
RELU	REctified Linear Unit
ResNet	Residual Network
R_u	Recommended (Overlooked learning)
SB	Sinus Bradycardia
SME	Subject Matter Expert
SmG	SmoothGrad
Snomed CT	Systematized Nomenclature of Medicine – Clinical Terms
SR	Sinus Rhythm
TAB	Tables
TF-IDF	Term Frequency - Inverse Document Frequency
TSE	Technical Self Efficacy

T_x	Explainability Score
Van	Vanilla Saliency
VGG-16	CNN named after Visual Geometry Group from Oxford
WFDB	Waveform Database
WMD	Weapons of Mass Destruction
XAI	Explainable AI

Chapter 1

Introduction

AI is becoming an invaluable tool in the decision making processes in healthcare. Quite often the safety risk associated with an AI's decision is low, this is definitely not the case in healthcare [2]. Human lives can be severely impacted by the decisions made by an AI. As a result the level of trust required in AI is extremely high. There are many challenges to achieving this high level of trust between the various stakeholders and the AI [3].

1.1 Problem Statement

The goal of this research is to gain insight into how deep learning models are making decisions for different user types and to quantify the features of importance. By improving the understanding of an AI model's decisions we can increase a user's trust in how the model identified features are contributing to outcome decisions.

To provide overall context this thesis will present a framework to define *Trust* and its individual constituent elements: *Explainability, Verifiability, Fairness, & Robustness* (Section 2.1.2). Within these elements of trust, "Explainability" will be the focal point of the exploration within this thesis. This body of work will examine how explainability ties into trust, and how this particular component of trust can be quantified for comparisons between different methods of explainability.

In the context of Artificial Intelligence, **explainability** refers to the act of explaining an AI's learned processes. The explanations of AI decisions are provided by Explainable AI (XAI) methods [4, 5].

Producing objective and quantifiable measurements of various explanations generated by different XAI methods allows these otherwise subjective explanations to be compared. Current literature provides some insights into smaller testable constituents of explainability:*Similarity, Stability, Novelty* (see Section 2.1). This thesis hypothesizes that by measuring these constituents of an explanation, it can be quantified and therefore compared with other explanations. This allows an objective framework to evaluate the quality of an XAI method.

1.2 Overview

This overview will first examine stakeholder attitudes toward AI, introduce a medical application for study, and present findings from a literature scoping review as a gap analysis.

1.2.1 Attitudes on AI & XAI

To explore the attitudes of clinicians and patients towards AI adoption we must first strive to understand the primary variables responsible for an individual adopting a position in favour of or against a specific technology. New models of user technology adoption suggest that the primary variable is technical self-efficacy (TSE) and it describes the level of comfort a user feels towards using a specific technology [6]. Other predictive variables include things like perceived usefulness and ease of use of technology, as well as perceived risk from a technology [6] [7]. Any exploration of attitudes toward AI adoption should consider these factors as integral towards developing an understanding of what steps need to be taken to improve user adoption.

Clinician Attitudes

A NHS funded study found that a majority of the clinical experts surveyed report never encountering any AI applications at work [8] see figure 1.1. Even with combining those clinicians who report encountering one or more AI applications at work, the overwhelming majority claim poor technological literacy, low levels of TSE, with fewer than 13% of the respondents being able to distinguish between 'Machine learning' and 'Deep Learning', see figure 1.2 [8].

While a sense of comfort with the use of AI tech maybe lacking among medical professionals, other variables like perceived usefulness are high and perceived risk is relatively low [8]. About 4 out of 5 medical professionals believe AI would be useful in their particular line of work, and almost everyone surveyed expresses no concerns with AI replacing their roles at work [8]. As a whole the majority of medical professionals express some concern regarding privacy laws especially with the roll-out of privatized AI applications [8] see figure 1.3. When medical doctors and other medical professionals were specifically asked about their level of fear towards AI, framed as AI being a bigger threat than WMDs, Medical doctors almost entirely rejected the concern that AI is a bigger threat than WMDs but the larger medical community sees AI as a significant threat see figure 1.4 [8].

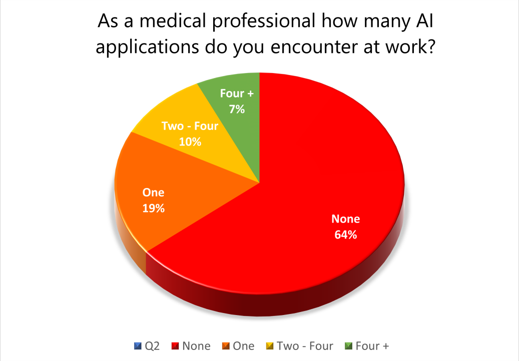


Figure 1.1: Clinician attitude - How many AI applications are encountered at work

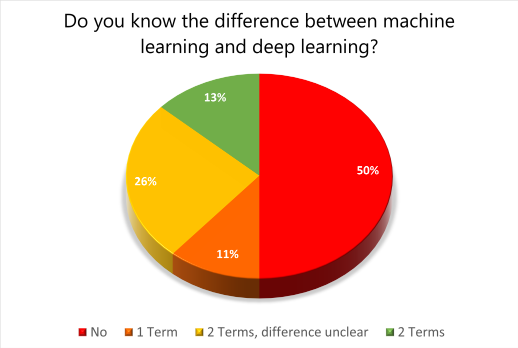


Figure 1.2: Clinician attitude - Knowledge about the differences in DL and ML

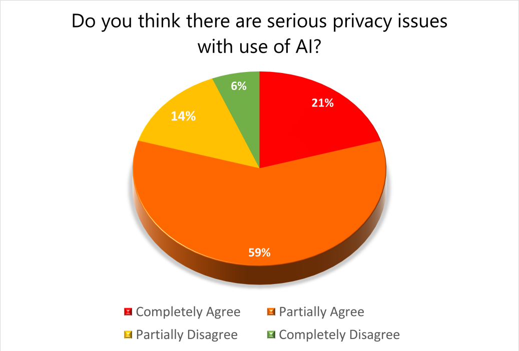


Figure 1.3: Clinician attitude - Personal position on privacy issues with AI

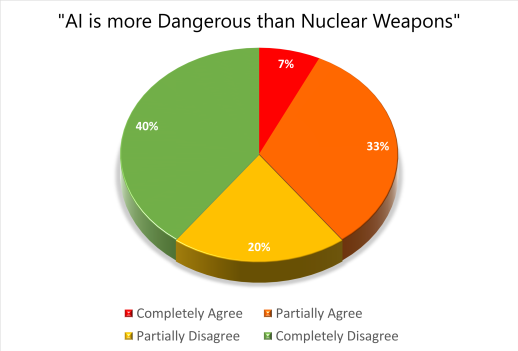


Figure 1.4: Clinician attitude - Do you Fear AI

The largest obstacle that remains between clinicians and full integration of AI applications in daily work is a sense of trust towards the technology. While this notion of trust will likely improve with improved TSE with AI. There are methods to improve the level of trust a clinician is willing to put in AI, by improving interpretability and explainability of AI decisions [3]. Most clinicians can recognize the benefits from advances in AI as they relate to more precise patient care and improved utilization of data generated worldwide [2] [8]. If the barriers to trust are going to be lowered, and clinicians made more willing to embrace the AI technology, efforts need to be focused on understanding trust and addressing the various components of trust.

General Public Attitudes

In the NHS survey [8] a small distinction between medical doctors and the larger medical community was made with the question regarding level of fear with AI. While medical doctors are relatively certain in not being afraid of AI, the larger community of people expressed a bit of fear see figure 1.4 [8]. In an Canadian Medical Association study of over 2000 respondents, conducted as a survey by Ipsos [1], a majority (57%) expressed an excitement about the future of AI used in healthcare. An overwhelming majority (77%) want to see more technological investments in healthcare, with 69% of the respondents believing that AI could address the challenges facing healthcare [1], see Figure 1.5.

When looking at the questions in the study [1], organized by views on utility, sense of trust in AI and concerns regarding AI, see figure 1.5, a pattern emerges. The general public is excited about the prospect of AI being of aid in healthcare and helping us address the challenges facing our modern health concerns. At the same time there is an obvious element of mistrust when physicians are out of the loop or privately managed healthcare services and health monitoring devices are powered by AI alone. There is a deeper sense of distrust when it comes to services and offerings by

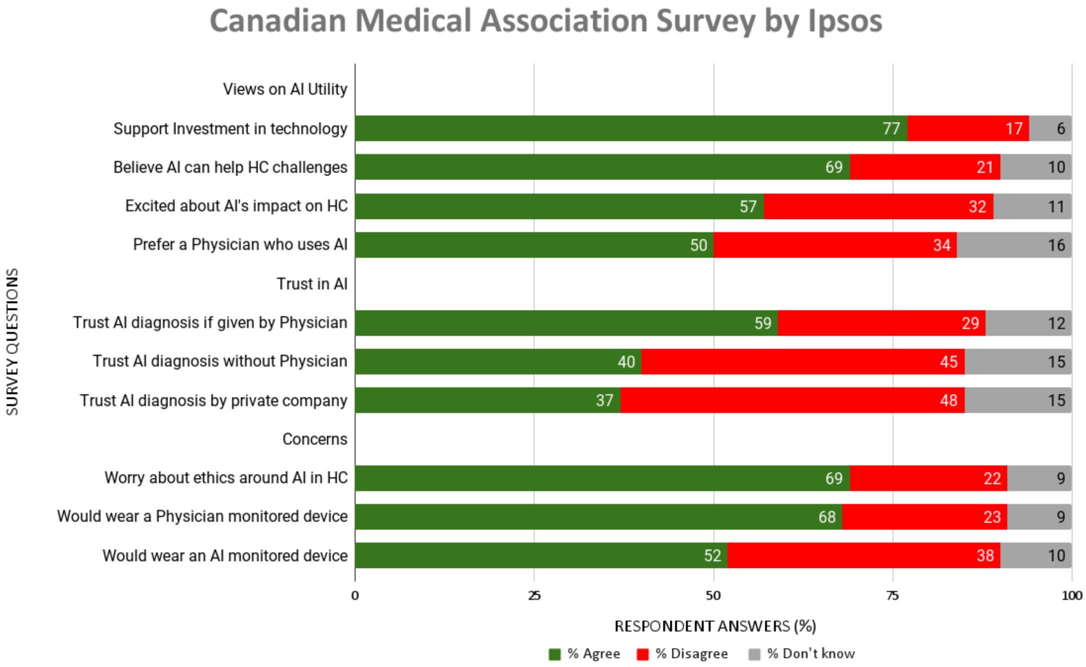


Figure 1.5: Summary excerpt of the CMA survey by Ipsos - 2018 [1] on attitudes of the Canadian public towards use of AI in Healthcare. *HC = Healthcare, AI = Artificial Intelligence.

the private sector in society that translates to a mistrust of the newer AI technology, especially if Physicians are kept out of the loop [1] [9]. There is an 'canonical' sense of trust that exists between the patient and their physician that has been built over a long history of patient-doctor interactions [9]. Part of this trust is grounded in the basis of expert knowledge on the part of the physician.

In reference to the use of AI in healthcare, to an average patient, the physician is expected to do more than just convey the findings of the AI. Rather a physician's role is seen as more of an intelligent expert user who can critically assess the recommendations of an AI and guide the patient so they may receive the best care possible [9]. This expectation on the part of the patients also means that there is an expectation by the physicians to be able to get insights from AI with some explanation and certainty [10] towards why those decisions are made. These relationships of trust between the AI, the physician, the patient and a few other stakeholders are explored in this thesis in the Section 2.1.

1.2.2 Healthcare Application

The healthcare application used in the process of this thesis is classification of a limited set of cardiovascular conditions using the ECG signal records from patients with 2-4 leads of data. The cardiovascular rhythms (conditions) included in this thesis are: Atrial Fibrillation (AFIB), Generalized Supraventricular Tachycardia (GSVT), Sinus Bradycardia (SB), and Sinus Rhythm (SR).

However, this research is applicable to any complex health data to which Deep Learning methods are applied.

1.2.3 Gap analysis

Benefits: Due to the volume of XAI generated explanations, especially as visualizations (image based explanations) there is a significant amount of coverage in literature of the utility XAI generated explanations offer to healthcare professionals [10]. There has been a lot of ground work done in determining and understanding the opinions of clinicians and lay-persons as they relate to AI models and explanations of said AI models used to justify the AI's decisions [2,8,10]. These works help identify the thinking patterns of stakeholders involved in using AI tools in healthcare, and help guide the direction of research for this thesis to identify what gaps in current knowledge and tools may exist.

Gaps: There are quite a few shortcomings in the current literature regarding XAI.

There are numerous definitions and taxonomies that result in a varied approach to the idea of generating explanations for an AI's black-box [11]. While having a multitude of approaches to elicit an explanation is advantageous as it provides a multi-dimensional view of the same processes, this raises the question of which explanation is objectively better, if any, and is there a way to compare the explanations with human performance and against other XAI generated explanations. While the idea of missing objective evaluation of explanations has been raised before [11], there are no known attempts to our knowledge that have aimed to functionally evaluate XAI generated explanations.

Aside from lacking any objective evaluation of the explanations generated by XAI, another gap is the lack of evaluation of the performance of the XAI methods in relation to clinical tasks they are sometimes used for [10]. Often when an XAI method is proposed there maybe some suggestion by the authors of how XAI method may aid in evaluating certain AI models on a specific clinical task [12, 13]. The amount of effort needed to explore the possible and useful XAI explanations for any given AI model for a specific clinical task is tremendously high.

1.3 Proposed solution

This thesis investigates if qualitative explanations of AI's learned processes can be quantified and objectively compared with other explanations. This investigation is performed using the learned processes of an AI classification model for cardiovascular rhythms.

This thesis:

1. Proposes a framework to examine any explanation using: **similarity, stability & novelty**.

2. Develops an AI model to perform classification of cardiovascular rhythms.
3. Generates a number of explanations from multiple XAI methods for the developed AI model.
4. Implements the proposed framework to quantify the qualitative explanations.
5. Tests the performance of explanations against each other as well as against the gold standard [human subject matter expert (SME)] explanations.

In performing the comparisons between explanations generated by various XAI methods and SME, investigating what remains *unlearned* by the AI is explored. This unlearned component is added to the framework in addition to similarity, stability & novelty.

1.4 Contribution

To the growing field of research in AI models in healthcare as well and XAI methods, this thesis contributes:

- A framework to evaluate trust in AI by examining qualitative explanations in quantifiable ways, based on the needs of stakeholders in the healthcare system.

1.5 Thesis Organization

The Figure 1.6 shows a visualization of a connected process: beginning with defining a healthcare application; identifying modelling question(s) and data type used for analysis; identification of relevant AI model(s); selection of appropriate explainable

AI (XAI) method(s), and finally quantifying the XAI's performance to improve the overall state of trust in the entire process.

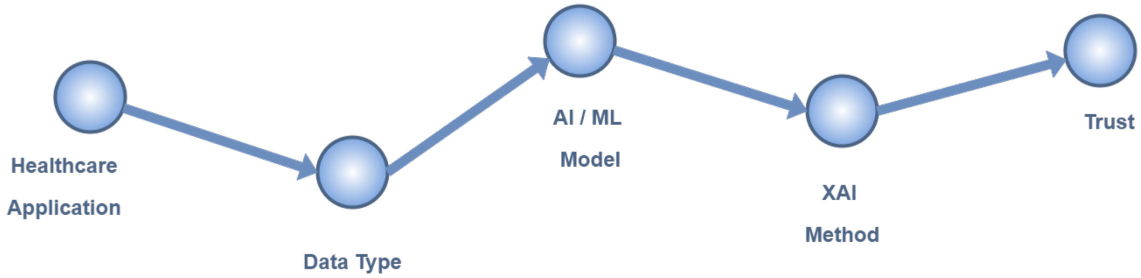


Figure 1.6: Visualization of overall journey of the thesis - Specifically dealing with the AI model in this section.

This thesis is organized as follows (components of the thesis are shown visually in Figure 1.6):

1. Literature Review - Presents a brief overview of the state of AI and XAI in contemporary research literature.
2. Domain Data - A detailed description of the data used in this thesis.
3. AI Model - Development and evaluation of the AI model that will be applied to the domain data and evaluated using XAI methods.
4. XAI Methodology - Selected XAI methods and how they are evaluated.

   Results - XAI generated explanations and results.

5. Quantifying Trust - Using trust metrics to evaluate the process.
6. Discussion - A discussion of results and their implications.
7. Conclusion & Future Directions - A summary of the thesis outcomes and next steps.

Chapter 2

Literature Review

This chapter discusses the current literature on the topic of trust and verification in AI. This literature review is performed in three stages: i) Identification stakeholders to understand concerns of users, ii) Elements of trust and verification of AI are extracted from literature to define parameters for the tools from the first stage, and iii) Algorithms and tools are identified for trust and verification of AI within the parameters as set by the second stage.

The goal of this review is to create a guide that consolidates information on state-of-the-art artificial intelligence (AI) & machine learning (ML) models, along with interpretable & explainable AI (XAI) methods and relates this information to various categories and disciplines in healthcare as shown in Figure 1.6.

In addition, the review will present a summary from literature of perceptions of trust towards AI. Finally, an equation framework to quantify trustworthiness of AI models used for problem solving in healthcare will be defined.

2.1 Trust in AI

As discussed in Section 1.2.1 there are a myriad of interactions and complex relations of trust between different types of stakeholders when it comes to the use of AI in healthcare. This section discusses the stakeholders, the elements of trust and how those elements of trust interact with the individual stakeholders affecting their levels of trust in the use of AI in healthcare.

2.1.1 Stakeholders

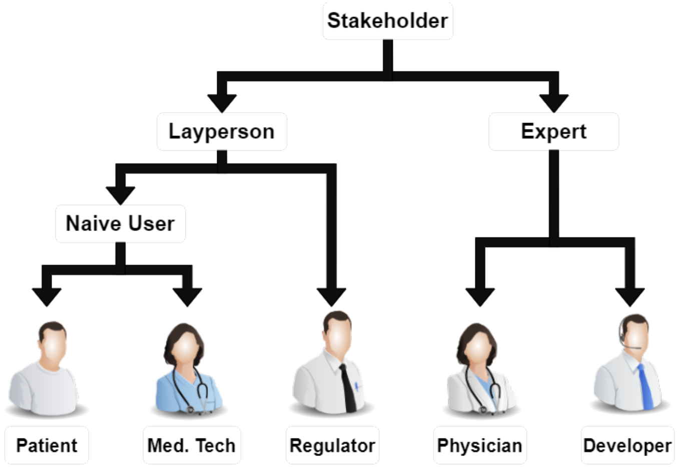


Figure 2.1: Visualization of various stakeholders in the use of AI in medicine

The first step is to identify and understand the stakeholders that are affected by the AI and its decision making. In Section 2.1.3 the individual concerns of of each stakeholder are explored in more detail. Based on broad evaluation of the literature [14–20] provide context for what a framework of AI systems and XAI methods need to provide users. Literature identified users, ranging from AI experts and subject matter experts to more general lay-persons; Each user type may need to interact with AI systems at different levels but may not posses the expertise required to analyse the performance of the AI or the AI's decisions. Lay-persons, as shown in Figure 2.3, range from regulators [14, 17, 19] (responsibility to ensure transparency in AI decisions) to more naive users (may only interact with the end output an AI decision making system produces). For example, these naive users may be the technical staff or employees [15,18,19] [16] who operate the machines or systems with the AI. The final naive user class is the client [14, 16, 17] who may only ever receive the final decisions made by the AI system and have little to no contact with the AI itself.

While the literature discusses the generalized case of stakeholder classes and their relative concerns. For the purposes of this thesis those stakeholder groups were translated to general groups that are more relevant to the healthcare industry. As shown in Figure 2.3 the experts are more distinctly defined to be Developers (AI experts) and Physicians (subject matter experts). While various types of academics (subject matter experts) may exist in the field of medicine, for simplicity only the physicians are mentioned on the figure. The naive users are also more granularly defined to be Medical technicians (technical staff or employees), and this category may also include administrators or nurses. Finally the last type of naive stakeholder in the healthcare industry is the patient (clients).

2.1.2 Elements of Trust

An important step in the path to improve trust for different types of stakeholders when it comes to the use of AI in healthcare, is to understand the word "trust" itself. The first step is to understand the general concepts of trust as they pertain to human interpersonal interactions. According to Ferrario et al. [21], trust between people exists in a variety of contexts and people have specific concerns they want addressed

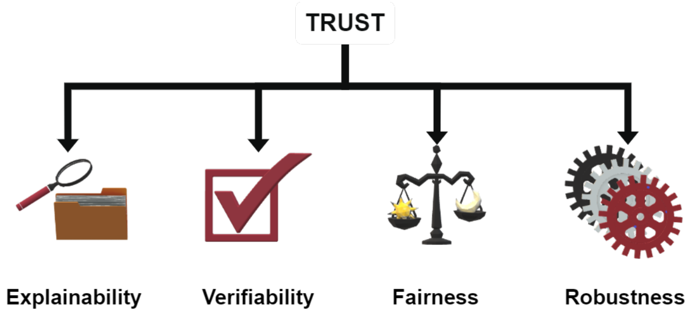


Figure 2.2: Elements of Trust
1

to achieve said trust, these concerns are discussed in more detail in Section 2.1.3. The “contexts“ within which trust is developed are concepts like *reliability* where if one party is able to demonstrate some consistency in how it performs a task, other parties learn to trust it in the context of reliability, even if it may not be trustworthy in other contexts [21, 22]. There are many other contexts like *transparency* and *fairness*, as well as being able to convey ones knowledge or *mastery* of a subject. These “contexts” of trust often have many overlapping definitions and are represented in a myriad of ways between literature [4, 5, 21]; Just as they apply to people they can be similarly applied to developing trust between humans and AI [21]. This thesis aims to reduce a large number of concepts within which trust is developed into four distinct elements, and discuss what overlap may exist between them (in Section 2.1.3.

As shown in Figure 2.2 the myriad concepts discussed in [4, 5, 21, 22] can be summarized into the four elements: Explainability, Verifiability, Fairness and Robustness.

Explainability allows an individual to comprehend what an AI has learned. This requires the AI being able to show its inner thoughts in its decision making process [4] [5]. This element of trust, explainability, is the prime focus this thesis as mentioned in the problem statement.

Verifiability allows the user to trust the decisions made by the AI because this element conveys that the current state of the AI has been validated by some expert [22] and the decisions it makes have some human oversight to ensure adequate rules and regulations for safety are followed [5].

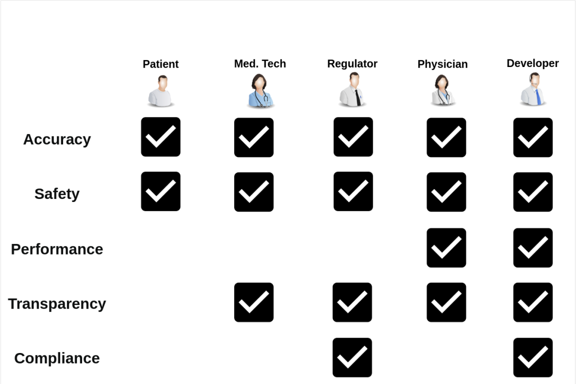
Fairness as an element of trust it's easy to grasp that an individual would be more trusting of an AI that makes fair decisions [4,5,14,20,21]. The trust in an AI's decisions by evaluating its fairness are affected by the individual need for correcting past inequities and the need for justice, which are prime reasons for why AI is even considered by many users as a new decision making paradigm in the first place [5]. Therefore, it is vitally important that an AI and its decisions be demonstrably fair to allow stakeholders to trust the AI's decision making.

Robustness deals with the concept of consistency in the AI's decision making especially as equipment, sensors, processors, and other hardware may change, as well as human subjects that data might be gathered from [4] [5] [16] [20]. This element of trust also conveys how much of a material improvement an AI produces in decision making in various different settings [4].

2.1.3 Impact of Elements of Trust

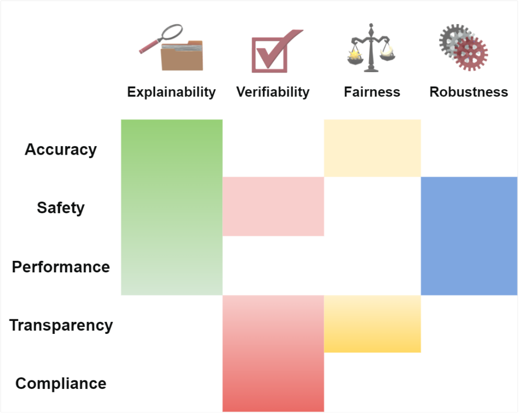
In the process of identifying the relevant stakeholders and understanding the elements of trust, the individual concerns of stakeholders to achieving trust become an important factor [4] [5] and they are what connect the stakeholders to the elements of trust.

The literature that was surveyed to develop an understanding of stakeholders and the elements of trust for this thesis, naturally provided the discussion that lead identifying five concerns that impact various stakeholders. These concerns are as shown in Figures 2.3 & 2.4; accuracy [14] [16] [20] [4] [5], safety [17] [20] [22] [4], performance [14] [16] [20] [4] [5], transparency [14] [17] [20] [4] [5], & compliance [14] [17] [19] [22].



	Patient	Med. Tech	Regulator	Physician	Developer
Accuracy	☑	☑	☑	☑	☑
Safety	☑	☑	☑	☑	☑
Performance				☑	☑
Transparency		☑	☑	☑	☑
Compliance			☑		☑

Figure 2.3: Stakeholders (columns) and their concerns (rows)



	Explainability	Verifiability	Fairness	Robustness
Accuracy	🟩		🟨	
Safety	🟩	🟥		🟦
Performance	🟩			🟦
Transparency		🟥	🟨	
Compliance		🟥		

Figure 2.4: Elements of Trust (columns) and interactions with stakeholder concerns (rows)

Accuracy refers to the general need for correct decisions made by the AI to have any semblance of trust [4, 5, 14, 16, 20]. This concern is common for all stakeholders, shown with check-marks in Figure 2.3. Accuracy can be demonstrated by an AI that is Explainable and Fair, shown with filled in coloured boxes in Figure 2.4.The two Figures 2.3 & 2.4 show that when it comes to trusting AI decisions, all stakeholders are concerned with the AI being accurate and that the need for accuracy means they are all concerned with the trust elements of Explainability and Fairness [4, 5].

Similar comparisons between all other concerns in Figures 2.3 & 2.4 can be made, to understand how stakeholders and the elements of trust are related.

Safety as a concern is about ensuring the behaviour of the AI remains consistent from one instance of implementation to another [4, 17, 20, 22]. That regulators can ensure the AI decision making meets standards that would apply to conventional machines or operators [17,20]. And that the decision making processes of the AI can be clearly explained when needed to ensure safe operation [20] [4].

Performance related concern is that AI needs to outperform a human or the conventional machine/tool in use currently, to justify the need for an AI based solution. This is a unique concern for only subject matter experts and AI experts [4,5,14], only the developers are concerned with making an AI that can compete with the current standards and subject matter experts are concerned with improving workflow. **Transparency** refers to the need to demonstrate that the AI's decision making is both fair and consistent. But the need for fairness here is distinct from the need to have an explainable AI [20]. The need for transparency exists for almost everyone who engages with the AI on some technical or operational level [4,5,20]. Technical users, regulators and experts want to ensure that all the decision making processes are testable and repeatable, something that comes with transparency. In literature explainable and transparent AI are used almost interchangeably [20], but this thesis separates the two ideas into two distinct axes to be able to better understand the distinctions between transparent AI and explainable AI and how each may relate to other elements and stakeholder concerns.

Compliance is a very narrow concern, but its very important to actually making an AI that can be trusted. Regulators and developers play an integral role in making sure an AI and its decisions remains trustworthy. To this effect, both of these stakeholders are concerned with ensuring that the AI systems being built remain compliant with various standards in the respective industries. This means that an AI has to be able to generate a verifiable audit trail of its decision making processes to be considered trustworthy [4, 5].

2.2 AI in Healthcare

2.2.1 AI Systems in Use

After understanding the point of view of stakeholders and their concerns regarding trust in AI. The next step is to understand on what is the AI doing and how its addressing problems in the field of Healthcare specifically.

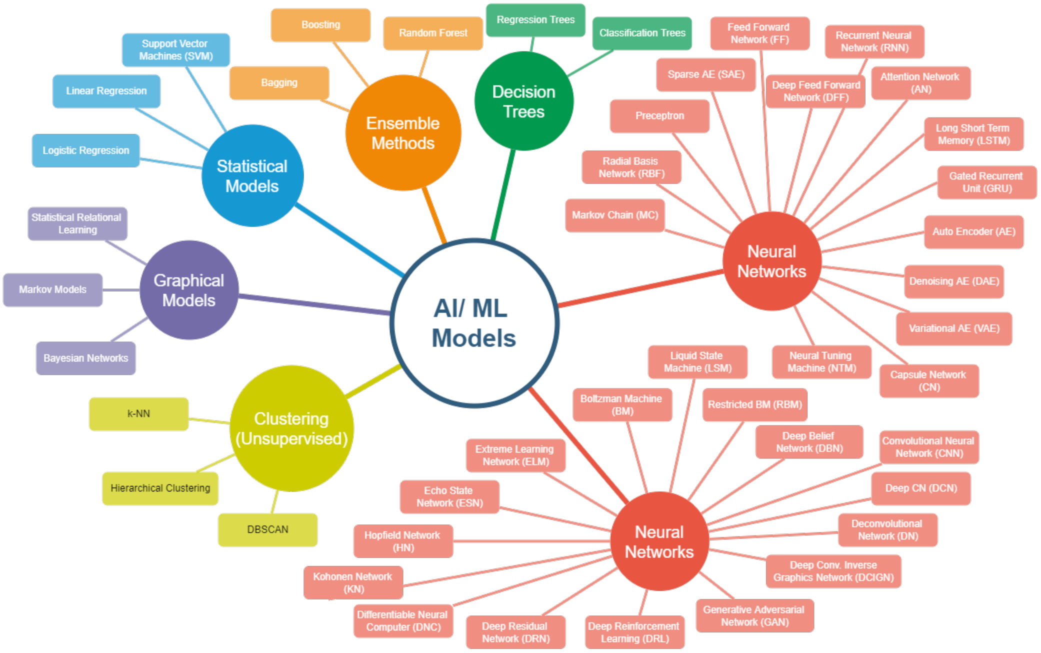

AI/ ML Models
Statistical Models
Support Vector Machines (SVM)
Linear Regression
Logistic Regression
Ensemble Methods
Boosting
Bagging
Random Forest
Decision Trees
Regression Trees
Classification Trees
Graphical Models
Statistical Relational Learning
Markov Models
Bayesian Networks
Clustering (Unsupervised)
k-NN
Hierarchical Clustering
DBSCAN
Neural Networks
Feed Forward Network (FF)
Recurrent Neural Network (RNN)
Sparse AE (SAE)
Deep Feed Forward Network (DFF)
Attention Network (AN)
Preceptron
Long Short Term Memory (LSTM)
Radial Basis Network (RBF)
Gated Recurrent Unit (GRU)
Markov Chain (MC)
Auto Encoder (AE)
Denoising AE (DAE)
Variational AE (VAE)
Liquid State Machine (LSM)
Neural Tuning Machine (NTM)
Capsule Network (CN)
Neural Networks
Boltzman Machine (BM)
Restricted BM (RBM)
Deep Belief Network (DBN)
Convolutional Neural Network (CNN)
Deep CN (DCN)
Extreme Learning Network (ELM)
Echo State Network (ESN)
Hopfield Network (HN)
Deconvolutional Network (DN)
Kohonen Network (KN)
Deep Conv. Inverse Graphics Network (DCIGN)
Differentiable Neural Computer (DNC)
Deep Residual Network (DRN)
Deep Reinforcement Learning (DRL)
Generative Adversarial Network (GAN)

Figure 2.5: All the Artificial Intelligence/ Machine Learning models

The AI hierarchy structure is adopted from DARPAs organization of relationships between AI algorithms [15]. Individual Neural Networks (NN) shown in Figure 2.5 are taken from the Asimov Institute's overview of NN taxonomies, which list state-of-the-art as well as historical networks, many of which are still in use [23]. Specific variants like U-Net (CNN), ResNet (DNN), BERT (AE) are not discussed because the scope for thorough coverage is too large & impractical for this review.

The Figure 2.5 aims to sort the myriad machine learning models into 6 total categories and the Neural Networks category is split into two separate groups; commonly used NNs vs uncommon NNs (top and bottom respectively). This separation on common vs uncommon use is based on whether a model provided any search results when searched in conjunction with Healthcare categories or healthcare disciplines shown in Tables 2.1 & 2.2 respectively in the pubmed database. Important to note that this separation does not suggest a conclusion that uncommon NNs could and/or would never be used for solving healthcare problems, rather an organizational artifact borne of a noted effect during literature search.

2.2.2 Utility of Various AI Types

Table 2.1: Categories of Health Applications with Relevant AI & Data Types in Literature

Categories of Care	AI	Data Type	Source
Robot-assisted Surgery	RNN	GRAPH	[24]
	LSTM	GRAPH	[24]
	GRU	GRAPH	[24]
	MCMC	GRAPH	[24]
	BN	TAB	[25]
	CNN	TIME SERIES	[26]
	CNN	TAB	[27]
	GAN	IMG	[28]
	SVM	TAB	[29]
	Logistic Regression	TAB	[29]
	k-NN	TAB	[29]
Virtual nursing Assistant	CNN, AE	TEXT	[30]
Administrative workflow	RF	TAB	[31]
	XGBoost	TAB	[31]
	AdaBoost	TAB	[31]
	Logistic Regression	TAB	[32]
	Logistic Regression	TIME SERIES	[33]
	FF	TIME SERIES	[33]
	BN	TIME SERIES	[33]
Fraud detection	Preceptron	TAB	[34]
	SVM	TAB	[34]
	Logistic Regression	TAB	[35]
	BN	TAB	[35]
	RF	TAB	[35]
	Classification Trees	TAB	[34]
	k-NN	TAB	[34]
	DBSCAN	TAB	[34]
Medication Management	FF	TAB	[36]
	RF	TAB	[37]
	Undisclosed (Private)	TAB	[38]
	Decision Trees	TAB	[39]
	RF	TAB	[39]
Clinical Trial Participation	DRL	TEXT, TAB	[40]
	AE	TEXT	[40]
	CNN	TAB, IMG	[41]
Cybersecurity	ANY	TAB	[42]
	ANY	TAB	[43]
Wearables and Monitoring	ML	TAB, TIME SERIES	[44]
Diagnositics	SEE TABLE 2.2		

Table 2.1 shows the various categories of healthcare and related AI models used in literature and identified by [45] as important categories for AI implementation in healthcare. Table 2.1 also shows the type of data used as inputs for the various AI models. The relevant references for each combination are also included in the table.

Table 2.2: Healthcare Disciplines with Related AI, Data types and Tasks in Literature

Healthcare Disciplines	AI	Data type	Task	Source
Allergy and Immunology	CNN	TAB	Using Deep learning to predict the syntax regulatory pathways used by the immune system	[46]
Anesthesiology	DL	TAB	Surgical decision making automation	[47]
Cardiology	CNN, DCN,ML	ANY	A review on diagnostic, and decision making AI available in the cardiovascular domain	[48]
Dentistry	ML	ANY	Improving reliability, reproducibility, accuracy and effectiveness in dentistry	[49]
Dermatology	DL Ensemble	IMG	Detecting and Assessing Melanoma in Skin Lesion Images	[50]
Diagnostic Radiology	CNN, DCN,ML	IMG	Review on Image segmentation and classification of tumors and other radiological findings	[51]
Endocrinology	Undisclosed (Private)	TAB	Optimizing insulin pump dosage compared to an expert physician	[38]
Emergency Medicine	DNN	TAB	Predicting clinical outcomes in emergency triage systems	[52]
Gastroenterology	RF	TAB	Predicting patients who will likely respond to a long term Crohn's treatment	[37]
Gerontology	RF	TAB	Analyzing gene expression to predict heart failure verified by patient records	[53]
Hematology	CNN	IMG	Diagnosing hematological diseases from blood smears	[54]
Hospice and Palliative Medicine	ML, NLP	TAB	Identify the need for and assist in faciliating end of life care	[55]
Internal Medicine / Family Med	DL	ANY	A review of the need to guide ethical development of AI for use by Primary Health Care providers	[56]
Medical Genetics and Genomics	RF, Decision Trees	TAB	Understanding the role of a gene mutation in TB drug resistance	[39]
Nephrology	CNN	TAB	Predict IgA Nephropathy	[57]
Neurology	CNN, RF	ANY	Diagnosing ischemic stroke and occlusions of blood vessels with AI and imaging	[58]
Nuclear Medicine	ML	ANY	Evaluating possible applications of AI in planning, scanning and interpreting in Nuclear Medicine	[59]
Obstetrics and Gynecology	ANN, CNN, ML	ANY	Review of papers presented at ASRM and ESHRE 2018 conferences on human reproduction	[60]
Ophthalmology	CNN, DCN	IMG	Primer on development of AI models for disease detection and diagnoses in opthamology	[61]
Pathology (anatomic / clinical)	CNN	IMG	Assisting with pathological screening of biopsy and histology images	[62]
Pediatrics	NLP, ML	TAB	Assisting physicians with clinical diagnoses by extracting clinically relevant data from EHRs	[63]
Physical Medicine and Rehabilitation	ML	TAB	Developing home rehab system for stroke survivors	[64]
Preventive Medicine	ML	ANY	Building infectious disease surveillence systems using big data	[65]
Psychiatry	FF	TAB	Determining doseage of psychiatric patients	[36]
Radiation Oncology	CNN, DCN	IMG	Tumor detection, segmentation, growth, drug dosage calculation	[66]
Rheumatology	AI	ANY	Modelling the risk of fragile fracture for patients with or at risk of osteoporosis	[67]
Sleep Medicine	DL, AE	ANY	Detection of sleep arousal to improve sleep studies	[68]
SURGERY			See Robot-Assisted Surgery - Table 2.1	
Urology	DL	TAB	Predicting prostate cancer with analysis of many molecular drivers	[69]

Table 2.2 shows the various types of healthcare disciplines compiled from the

AAMC (Association of American Medical Colleges) and related AI models, as well as the type of data used and tasks that were performed with the AI. The relevant references for each combination are also included in the table.

2.2.3 Model Parameters

There are many adjustable model training parameters (different from learnable parameters in an AI - see section 4.1) known as hyperparameters that can be manipulated to optimize the performance of the CNN (AI).

Hyperparameters that were used in the model in section 4.1 include:

1. **Freezing/Unfreezing Training** - Disabling or enabling the training (changing) of the trainable parameters.

2. **Loss function** - An error function that is minimized to find the optimal solution, and to understand how the model is performing at any given time [70].

3. **Optimizer** - Used to adjust weights in the model to minimize the loss function [71].

4. **Learning Rate** - How much the weights are modified during training [71].

5. **Momentum** - Helps determine the rate of acceleration for gradient descent, it adds a lever of control in adjusting the learning rate [72].

6. **Test, Train, Validation Split** - The percentage split of the data between training, validation and testing.

7. **Batch Size** - The number of samples presented to the AI model at one time for training.

8. **Learning Epochs** - One iteration of presenting all the training/testing data, one epoch may require multiple batches.

7. **Early Stopping Parameters** - Stopping the training earlier than the predefined number of epochs based on the performance of the AI model monitored at the end of each epoch [71].

2.3 XAI Methods in Use

2.3.1 Types of XAI methods

Table 2.3: AI models, transparency parameters (Inherent or Post Hoc) and Data Type, with related and compatible XAI methods and related properties of the XAI method

AI Models	Tran.	AI Data Type	XAI Methods	Vis. Type	LO/GL	MA/MS	Source
Logistic Regression	IN		-	-	-	-	-
Linear Regression	IN		-	-	-	-	-
Decision Trees	IN		-	-	-	-	-
K-NN	IN		-	-	-	-	-
Rule Based Learners	IN		-	-	-	-	-
General Additive Model	IN	TAB	GA2M	TAB	GL	MS	[73]
Bayesian Model	IN	GRAPH, TAB, TEXT	iBCM	GRAPH, TAB,TEXT	GL	MS	[74]
Ensemble Methods & Decision Trees	PH	TAB	defragTrees	TAB	GL	MS	[75]
Ensemble Methods & Decision Trees	PH	TAB	inTrees	TAB	GL	MS	[76]
Random Forest	PH	TAB	Forest Floor	TAB, IMG	Both	MS	[77]
CNN, RNN, LSTM, GRU	PH	IMG, TAB	SWAF (Stacking With Auxiliary Features)	TEXT	LO	MS	[78]
SVM	PH	TAB	hybrid Rule-Extraction	TAB	GL	MS	[79]
SVM	PH	TAB	Bayesian Method	TAB	GL	MS	[80]
SVM	PH	IMG	Contribution Propagation	IMG	LO	MS	[81]
Linear SVM	PH	IMG, TAB	Heatmap coloring viewer	TAB, IMG	GL	MS	[82]
NN, ANN, FF	PH	TAB	RxREN (Rule Extraction by Reverese Engineering)	TAB	GL	MS	[83]
NN, Perceptron, RNN w/ GRU,	PH	TAB	Tree Regularization	TAB	GL	MS	[84]
NN, DCN,	PH	IMG	Distilling ensemble	IMG	GL	MS	[85]
...	...	...	...	...	...	...	...
CNN, DCN	PH	IMG	SG (SmoothGrad)	IMG	LO	MA	[86]

The Table 2.3 presents a combination of AI model types, identifies if the models are transparent inherently (IN) vs. need post hoc (PH) transparency methods, as well as the type of data used by the AI model. This information on AI models is combined with the information on related and/or compatible XAI methods, their visualization of the output data; which includes Tables (TAB), Graphs, Text outputs, and Images (IMG). The scale of the XAI evaluation method is presented as well, whether the XAI evaluates the AI model's Global (GL) properties or Local (LO) properties. If the entirety of the AI model is evaluated then its a global scale XAI vs. if a finite portion of the AI model is evaluated by the XAI, then its a local scale XAI model. Lastly, the XAI models specificity is presented in the table. If the XAI model can work with any AI model then its Model Agnostic (MA) vs. if it only works with specific AI models then its Model Specific (MS) [16] [20] [17].

Table 2.3 is only a truncated summary of the entire list compiled for this literature review, the entire complete table is present in Appendix A in Table A.2.

2.3.2 Utility with various XAI methods

The following selection of XAI methods are chosen for the purpose of detailed evaluation in this thesis: Vanilla Saliency, SmoothGrad, GradCAM, & GradCAM++. More details about the implementation of these methods is discussed in Section 5.1. A brief primer to understand the general concepts of each XAI method are presented here.

Vanilla Saliency was designed to address the need of visualising the neuronal activity of Deep Convolutional Networks (DCNs). For an arbitrary number of layers in a DCN, compute the gradient of output relative to the input:

$$\frac{\partial Output}{\partial Input} \tag{2.3.1}$$

Any pixels with positive gradient values can be identified as pixels where the output value will increase if the input value is increased, this is done by using a RELU activation function in the back-propagation process. Through back-propagation it identifies all the pixels in an input image that when perturbed will create the most change in the output image. This can be done for any layer (not necessarily the final output layer). Finally, the process is to map out a visualization (saliency map) by highlighting only the most informative pixels on the input image for any given classification [87].

SmoothGrad is a more coherent mapping method than Vanilla Saliency. In the previous method, there are noisy pixels with gradients high enough to end up appearing as important input pixels. SmoothGrad, adds (to each pixel) stochastic Gaussian noise in the otherwise same formula that was used in Vanilla Saliency. And then the final gradients are capped at a very high value to capture most of the relevant gradients and remove the errant 'noise' in the gradient mapping. This process then goes one step further and multiplies the final gradient map with the original image pixel values to produce a very coherent and sharp final visualization [86].

GradCAM (Gradient-weighted Class Activation Mapping) produces a coarse visualization highlighting the areas on the input image that are important regions for the classification decision by the AI model. To produce the GradCAM visualization, the model's convolutional layers are cut-off (up to whatever layer the user wishes to analyse) and the fully connected classifier layers are attached up to that cut-off convolutional layer. The input image is passed to the newly modified model and once again a gradient of the classifier's output layer (before the softmax activation layer) relative to the feature maps of the convolutional layer:

$$weight = \frac{\partial OutputClassScore}{\partial Conv.LayerFeatureMap} \tag{2.3.2}$$

is determined via back-propagation. This weight is multiplied by the activation map and passed to a RELU activation function in the forward direction to create the GradCAM mapping. Shown here:

$$GradCAMHeatMap = RELU(weight * conv.layerfeaturemap) \tag{2.3.3}$$

This heatmap is then overlayed onto the original input image and provided to the user to visually identify the AI's understanding of the areas of high importance in the input image [88].

GradCAM++ is a more sophisticated version of GradCAM. The weight calculated during back-propagation in GradCAM is modified:

$$weight = ClassGradientWeights * RELU(\frac{\partial OutputClassScore}{\partial Conv.LayerFeatureMap}) \tag{2.3.4}$$

followed by a forward direction creation of the Heat Map. The ultimate result of this method is the ability to produce better visual localization of objects (in a multi-class problem) in a single image [89].

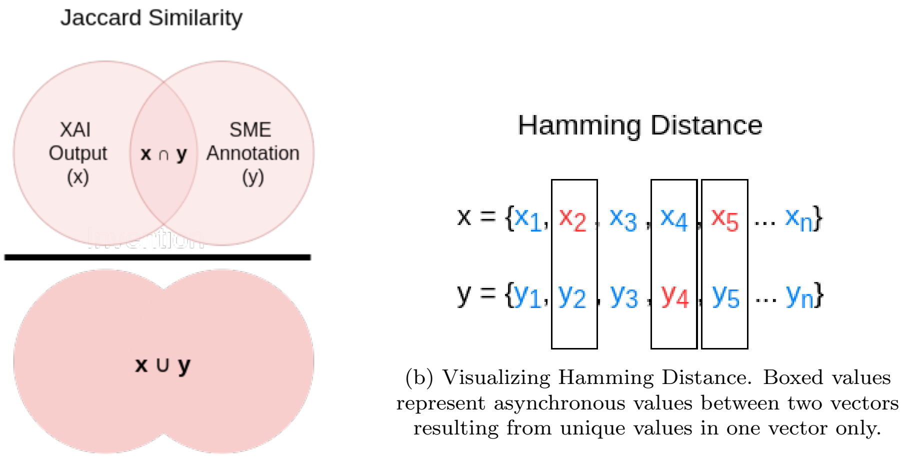
2.3.3 How XAI impacts people

As presented in Section 2.1.3 and Figures 2.3 & 2.4 everyone is affected by the need for explainability when it comes to developing trust in an AI's decision making. And that explainability element that will adequately address trust for a user will do so by providing insights into the AI's **accuracy**, provide the user with some assurance that the AI's decision making is **safe** and provide some **performance** related advantage for some users.

This thesis uses the Explainable AI (XAI) methods presented in Section 2.3.2 and evaluates them for their ability to provide a measurement of accuracy of the underlying AI. Measures the XAI's performance for consistency, this provides a measurement of safety of the underlying AI. This thesis also evaluates the XAI methods for any insights on AI model's novel learned processes to test if the performance of the workflow is affected by the underlying AI.

Accuracy is more finely defined as the _(similarity) between XAI and SME's outputs. Safety is evaluated by the _(stability) of the XAI relative to other XAI methods. Lastly, the XAI methods evaluate the performance of the AI by identifying any _(novelty) in learning or_(overlooked learning) by the AI.

Similarity is first comparison that provides a measure of objectivity in an XAI's trustworthiness. This is done with comparing the XAI method's output, which is its determination of the AI's learning, against the SME annotations. The two distance metrics used to determine the similarity between the pairwise array are Jaccard Similarity and Hamming Similarity.



(a) Visualizing Jaccard Similarity equation as a Venn diagram. The intersection of the two datasets divided by the union of two datasets.

(b) Visualizing Hamming Distance. Boxed values represent asynchronous values between two vectors resulting from unique values in one vector only.

Figure 2.6: Visualizing Similarity with Jaccard Similarity and Hamming Distance Metrics

$$D(x,y) \ = \ 1 \ - \ \frac{|x \ \cap \ y|}{|x \ \cup \ y|} \tag{2.3.5}$$

Jaccard Similarity shown above in equation 2.3.5, is commonly used as a predictor of similarity between two datasets [90]. It is also used in machine vision applications to predict labels for image segments [91]. In this thesis, the Jaccard similarity metric is used to identify the similarity between the pairwise arrays of XAI outputs and SME annotations. Figure 2.6a shows of visualization of the equation 2.3.5, the Jaccard similarity metric is derived by dividing the intersection of the two datasets by the union (combined total minus the intersection). An metric value of 1 represents that the two datasets are identical, and a value of 0 represents no overlap at all.

$$D(x,y) \ = \ 1 - \frac{1}{n} \sum_{i=1}^{n} 1_{xi \ \neq \ yi} \tag{2.3.6}$$

Hamming Similarity shown in equation 2.3.6, is the inverse of the normally used

Hamming distance metric. Hamming distance provides a metric score representing all the instances of dissimilarity between two datasets. The Hamming distance visualization in Figure 2.6b shows that a unique cluster identified by the XAI output only or an annotation identified by the SME only, will be identified as an asynchronous event between the two vectors and reduce the similarity metric. A Hamming similarity metric of 1 represents two identical datasets, and 0 represents no overlap at all.

Box-and-whisker (BW) plots are used to show how accurately each XAI method can present the underlying learning of the AI model. Given a correct classification by the AI, with no overlooked features and no novel learning possible. The best XAI method output has a similarity score of 1.0 for Jaccard or Hamming similarity metrics. Similarly, for an incorrect classification by the AI the best XAI method identifies a lot of differences between the SME annotations and the XAI output. Thus giving a score of 0.0 for Jaccard and Hamming similarity metrics. Therefore, a BW plot helps visualize the performance of the XAI method, where incorrect classifications by the AI are all in the lower inter-quartile range of the BW plot.

The 'box' shows where 25th, median (50th) and 75th percentile (the inter-quartile range(IQR)) of the underlying data lie on the y-axis (the similarity metric). The upper and lower 'whiskers' represent the min and max points of data that are 1.5 times beyond the IQR in either direction, points beyond the 1.5 IQR are outliers. 'Outlier' is only a statistical term to represent the data point's distance from the IQR. It may or may not have any real world significance, beyond representing the skewness and/or kurtosis of the dataset.

Stability is the next comparison that provides a measure of consistency of an XAI's output. Measuring the consistency of an XAI method presenting the underlying

AI learned processes affect a user's trust in the underlying AI's safety. This measure is provided by comparing the XAI method's output to all other XAI methods' outputs for every test sample.

Pearson correlation, shown in Figure 2.7, determines the degree of correlation between the two datasets. The Pearson correlation coefficient provides a measure of distance between the two datasets, and also a measure of directionality of that relationship. This directionality is not important for the purpose of measuring the stability of the XAI methods.

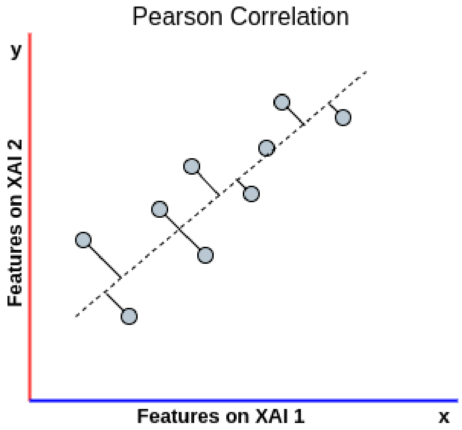


Figure 2.7: Visualizing Pearson Correlation - Stability Metric. Determines the relationship between two datasets of values [outputs of XAIs].

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}} \tag{2.3.7}$$

Pearson correlation output is between -1 and 1 and is a true metric or distance measure [90] because it meets the triangular inequality required for a distance measure. A metric that is not contained between 0 and 1 instead of the -1 and 1, makes comparisons to other metrics used in this thesis difficult. Thus the values of r from Equation 2.3.7 are remapped onto a space between 0 and 1 in Equation 2.3.8 by taking the absolute value.

$$r = abs(\frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}}) \tag{2.3.8}$$

where:

r is the Pearson Correlation coefficient.

x_i is value of a single feature in the first XAI output,

$\overline{x}$ refers to the mean of values in the the first XAI output,

y_i is value of a single feature in the second XAI output,

$\overline{y}$ refers to the mean of values in the the second XAI output,

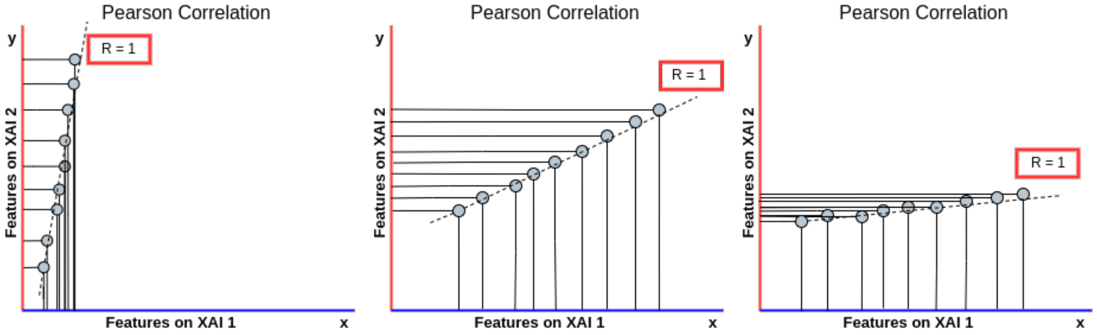


Figure 2.8: Pearson Correlation coefficient (R) is independent of slope, it is only a measure of the relationship between two datasets.

It is important to note that the Pearson Correlation coefficient is not a measure of steepness of the slope as shown in figure 2.8, rather it is a measure of how closely data from two variable (XAI methods) maps onto a straight line. There is no assumption of direct relationship between the two XAI methods, rather a confounding variable (learning by the underlying AI) can be the source of correlation between the two variables (XAI methods).

Novelty and **Overlooked** learning is the last XAI comparison. This comparison provides insights about the performance of the underlying AI method. These insights are acquired by evaluating individual features highlighted or overlooked by the XAI output relative to the SME annotations. In this thesis, the novel features and overlooked features are defined as:

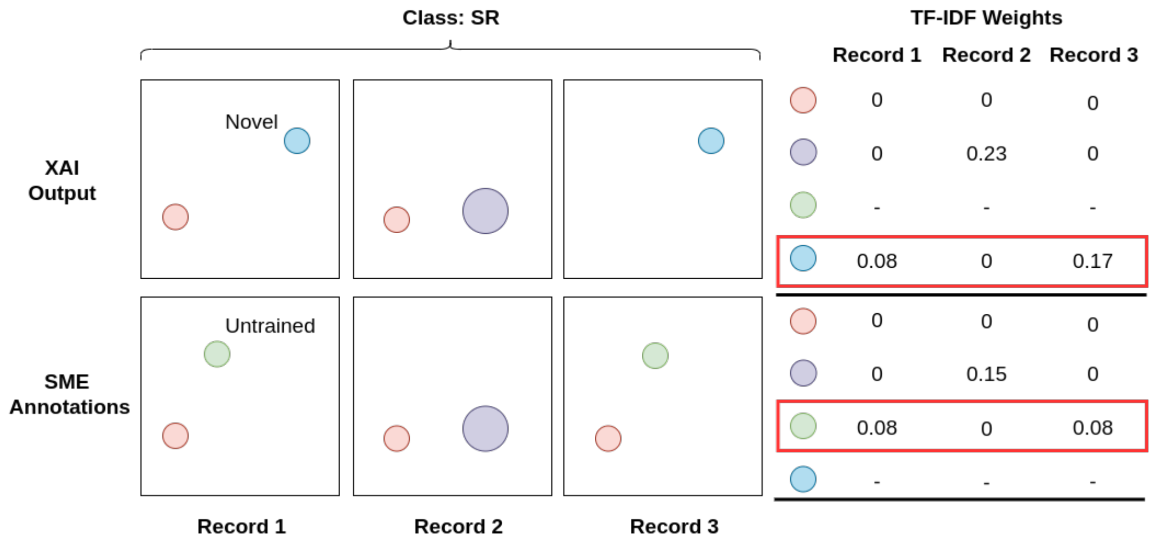
Novel features are individual features (combinations of pixels) in an ECG record identified in an XAI output as important for classification, but not identified by the SME.

Overlooked features are individual features (combinations of pixels) in an ECG record that correspond to SME annotations but not identified by the XAI output.

Term Frequency - Inverse Document Frequency (TF-IDF) method is used in this thesis to determine the predictive value of novel features and/or overlooked features towards classification of an ECG record. This method is commonly used in literature for text mining, to determine the semantic importance of terms (features) in a document (ECG record) relative to all other terms in that document within a collection of related documents [91]. In this thesis, TF-IDF is applied to collections of ECG records grouped by their classification.

TF-IDF method first creates a vocabulary of all features present in ECG records of a single class. As shown in equations 2.3.9 & 2.3.10, TF term identifies the frequency with which a feature is present in an ECG record. IDF term identifies the semantic importance of the feature to the ECG record, giving a weighting for the feature relative to the vocabulary.

Figure 2.9 is a cartoon representation of ECG records, to demonstrate how TF-IDF can identify the importance of a feature to a single ECG record. In the thesis, any features common between the SME annotation and XAI output are excluded from this analysis. Thus, the only features being evaluated in figure 2.9 for their importance are those that are already either completely novel (blue) features or completely overlooked (green) features. The remaining features (red and purple) in figure 2.9 are ignored because they are common between XAI output and SME annotations. Calculation of TF-IDF weights shown in figure 2.9 is made using equations 2.3.9 and 2.3.10.



	Record 1	Record 2	Record 3
🔴	0	0	0
🟣	0	0.23	0
🟢	-	-	-
🔵	0.08	0	0.17
🔴	0	0	0
🟣	0	0.15	0
🟢	0.08	0	0.08
🔵	-	-	-

Figure 2.9: Visualization of TF-IDF used to determine the significance of features identified by an XAI output and SME annotations for a given record. Each record is an ECG test sample, coloured circles represent features, blue and green features are novel and overlooked features respectively.

Equation 2.3.9 shows how the TF-IDF score for novel features is determined. Equation 2.3.10 shows how the TF-IDF score for overlooked features is determined. Note t_o value in equation 2.3.10 is calculated differently than t_n in equation 2.3.9.

$$\begin{aligned} TF-IDF_n &= TF \;.\; IDF \\ TF &= log(1+freq\ (t,d)) \\ IDF &= log\ \left(\frac{N}{count\ (d \in D:\ t_n\ \in d)}\right) \\ t_n &= (X - \ (A \cap X))\ \ ;\ \ X = [x_1, ...,\ x_i]\ ,\ A = [a_1, ...,\ a_j] \\ d &= [t_n|\ y_{pred}\ \cap\ y_{true}] \subset D\ :\ D = (A \cup X) \end{aligned} \tag{2.3.9}$$

where:

t_n is a single term (feature) from X not present in A,

d is a single document (ECG record),

D group of all documents (all records for a single class),

X refers to XAI output array,

A refers to SME annotations array.

$$\begin{aligned} TF-IDF_n &= TF \;.\; IDF \\ TF &= log(1+freq\ (t,d)) \\ IDF &= log\ \left(\frac{N}{count\ (d \in D:\ t_o\ \in d)}\right) \\ t_o &= (A - \ (A \cap X))\ \ ;\ \ X = [x_1, ...,\ x_i]\ ,\ A = [a_1, ...,\ a_j] \\ d &= [t_o|\ y_{pred}\ \cap\ y_{true}] \subset D\ :\ D = (A \cup X) \end{aligned} \tag{2.3.10}$$

where:

t_o is a single term (feature) from A not present in X.

All else is the same as equation 2.3.9

2.3.4 Limitations of XAI

Most reviews and primary research papers in literature that deal with trust in AI or explainable AI try and define the concepts of explainable AI and differentiate them from concepts like interpretability and transparency etc. [17] [92]. Then there are papers that actually suggest methods for evaluating XAI with metrics, but they mostly end up creating some new XAI methodology that is easier to evaluate than the current methods and objective XAI evaluations is not actually accomplished [93]. There are fewer papers still discussing the need to build trust especially in the context of AI and healthcare [9] [10].

With the slight exceptions of [88] & [89] there aren't any papers that present a tangible method of objectively evaluating the performance of an XAI method. Even the two papers [88] [89] make a very limited case of 'objective' evaluation which requires a human subject to ultimately determine if the interpretable XAI is indeed interpretable. Also the arguments for trust are very limited and the scope is focused on determining if the XAI methods Grad-CAM and Grad-CAM++ are objectively and subjectively good. To determine if the XAI method is objectively good the XAI has to be 'interpretable' and 'faithful'.

2.4 Quantifying Trust

Based on the literature review regarding the need to improve user trust in AI, the important concepts identified in building trust are combined into a series of equations.

This will help provide an objective metric or numerical value of trustworthiness to an AI model. Similar to the mathematics of trust [94] and algorithms used to convey the concept of trust in human (adversarial and non-adversarial) interactions [95] and human to machine interactions [96] the following equation is presented by the author as a method to capture and quantify important elements of trust.

$$T = \left(\left[\frac{\sum_{i} (w_x T_x)_i}{\sum_{i} w_i} \right] + w_v T_v + w_r T_r + w_f T_f \right) / \sum_{n} w_n \tag{2.4.1}$$

Equation 2.4.1 is the full Trust equation, combining all the elements of trust; Explainability (T_x), Verifiability (T_v), Robustness (T_r) and Fairness (T_f) and the corresponding weights (w) that will have to be determined experimentally.

$$E_{sim} = \frac{\sum(w_0 M_0, w_1 M_1, ..., w_n M_n)}{\sum w_n} , \ 0 \leq M \leq 1 \tag{2.4.2}$$

$$E_{stab} = \frac{\sum(w_0 N_0, w_1 N_1, ..., w_n N_n)}{\sum w_n} , \ 0 \leq N \leq 1 \tag{2.4.3}$$

Equations 2.4.2 and 2.4.3 with names and weights of test metrics:

$$E_{sim} = \frac{\sum(0.5 M_{Jaccard}, 0.5 M_{Hamming})}{1} , \ 0 \leq M \leq 1 \tag{2.4.4}$$

$$E_{stab} = \frac{\sum(N_{PearsonCorr})}{1} , \ 0 \leq N \leq 1 \tag{2.4.5}$$

E_{sim} in equation 2.4.2 combines individual metrics (M) used to evaluate the XAI methods and their respective weights (w), represented by $w_n M_n$. Jaccard and Hamming similarity metrics and both are true metrics (distance measures) [90] such that the values of M_n are contained between and 0 and 1.

E_{stab} in equation 2.4.3 presents the individual metrics (N) used to evaluate the XAI method's stability and their respective weights (w), represented by $w_n N_n$. Pearson Correlation is the stability metric true metrics (distance measures) [90] such that the values of M_n are contained between and 0 and 1.

$$T_x \;=\; \frac{\sum(\; w_{sim}E_{sim},\;\; w_{stab}E_{stab}\;)}{\sum w} \tag{2.4.6}$$

The Equations 2.4.4 & 2.4.5 relate the concept of similarity and stability to the explainability element of Trust shown in equation 2.4.6. Equation 2.4.6 is the combination of similarity and stability equations that combine together here to provide one trust measure that combines into the overall trust equation 2.4.1. Each additional XAI method used to evaluate the AI model, adds an additional T_x term to the equation 2.4.1.

$$K_n \;=\; \frac{\Sigma(Record_n|XAI_{novelty})}{\Sigma(Record_n)} \tag{2.4.7}$$

$$R_u \;=\; \frac{\Sigma(Record_u|XAI_{overlooked})}{\Sigma(Record_u)} \tag{2.4.8}$$

The Equations 2.4.7 & 2.4.8 are used in equation 2.4.9. K_n and R_u provide a measure of how much novel learning or overlooked learning was identified by a single XAI method respectively.

$$E_{XAI} \;=\; [T_x\;,\; K_n\;,\; R_u\;]\;\;,\;0 \leq T, K, R \leq 1 \tag{2.4.9}$$

Equation 2.4.9 is a vector of values that provides insight on the utility of the individual XAI method from the perspective of Explainability as shown in Figure 2.4. T_x refers to trust conferred on to the individual XAI method derived by Equation 2.4.6. K_n refers to any novel Knowledge that the AI model may have learned that the XAI method can identify. A value approaching 1 means each record had a novel feature identified. R_u refers to any Recommended overlooked learning by the AI model that the XAI can identify, a value approaching 0 means no overlooked features were identified. This vector of values is not meant to combine with any other equation, rather be a standalone vector to inform the user on the utility of the evaluated XAI.

Chapter 3

Domain Data

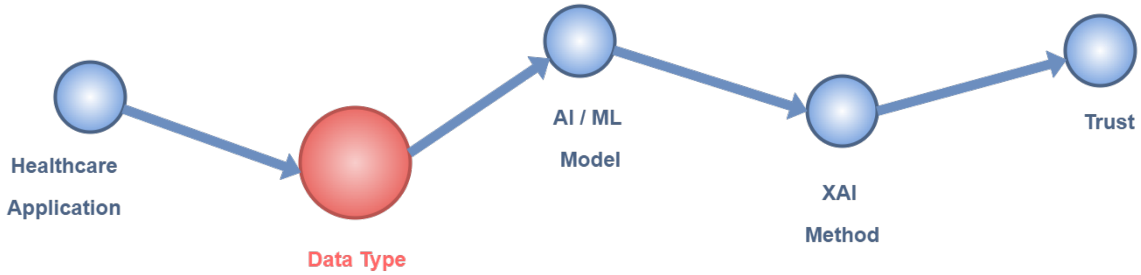


Figure 3.1: Visualization of overall Journey of the Thesis - Specifically dealing with the Task and Data type in this section.

To evaluate XAI methods it was necessary to apply a machine learning model to a suitably complex data set. The data and its preparation method used for training and testing the machine learning model are described in this section.

3.1 ML Objective: ECG Classification

This thesis utilized continuous time-series data to evaluate ECG signals from patients. The time-series data was used to train a classifier model to identify four cardiovascular conditions.

3.1.1 Domain and Model Datasets

Three databases were used for this Thesis, two ECG databases were used and one machine vision database. ImageNet was the machine vision database used as the default pre-trained weights for Keras' implementation of VGG-16 [71] & [97]. The imageNet dataset has 1000 classes of images, and contains 1,281,167 images for training, 50,000 images for validation and 100,000 images for testing.

The 'MITBIH' database is an open access database acquired from Physionet repository of physiological signals, officially referred to as 'MIT-BIH Arrhythmia Database'[1] . 'Chapman' database is also an open access database produced and maintained by the Chapman University and Shaoxing People's Hospital in China. Officially referred to as 'Chapman-Shaoxing database' [2] [98]. The two ECG databases are discussed in detail in Table 3.1.

[1] https://physionet.org/content/mitdb/1.0.0/

[2] https://figshare.com/collections/ChapmanECG/4560497/1

Table 3.1: Statistics of ECG signal databases [Chapman & MITBIH] used for training, as original datasets and after modification with reduced classes and harmonized record lengths

Database	**Original**						
	Age Range	**# of Subjects**	**# of Records**	**Record Length**	**Sample Freq.**	**Classes**	**Leads**
Chapman	4-98	10646	10646	10 sec	500 Hz	11	12
MITBIH	23-89	47	48	30:06 min	360 Hz	15	2
	Modified						
	Used in Training	**# of Subjects**	**# of Records**	**Record Length**	**Sample Freq.**	**Classes**	**Leads**
Chapman	Step 1	10646	10646	10 sec	500 Hz	4	4
MITBIH	Step 2	47	7414	10 sec	500 Hz	4	2

3.2 Data Labelling

The two ECG datasets presented below are utilized by the CNN model, as presented in Section 4.1 to make classification decisions between four distinct cardiovascular rhythms.

Chapman Dataset: This dataset has a total of 10,646 records of 12 lead data. Each record corresponds to an individual subject (patient), and is 10 seconds in length, sampled at 500 Hz. The original ECG data was classified by 11 different rhythms, these 11 rhythms were then combined together to form 4 groups of rhythms. These mergers of rhythm classes were done by trained Cardiologists as presented in [99].

MITBIH Dataset: This dataset originally contains 2 lead ECG recordings from 47 subjects with 30:06 minute long ECG recordings for each individual subject, sampled at 360 Hz [100] & [101]. The original data was classified by 15 different rhythms, these were once again grouped into 4 groups to reduce the number of total classes. The merging of classes in the MITBIH dataset was done by combining classes with the same SNOMED CT codes as the 11 classes in Chapman dataset. The SNOMED CT codes for the classifications from both datasets were acquired from the PhysioNet competition [17]. In addition to the overlapping SNOMED CT codes, an online ECG library maintained by Dr. Robert Buttner of Monash University and Alfred Emergency & Trauma Center - an expert in ECG and Ultrasound diagnostics, was used to confirm if some labels under the MITBIH dataset could be merged into the GSVT category. If no unambiguously clear answer was found by either means (SNOMED CT and ECG library) the label was placed into a separate category and omitted from consideration. A full table of all conditions and their corresponding SNOMED CT codes are presented in the Appendix A

Dataset Overlap: The Chapman dataset provides 12 lead ECG data for each of the 10,646 records of data. The MITBIH provides 2 leads of ECG data for each record that are various combinations of the 4 ECG leads (II, V1, V2, V5). Four leads of data (II, V1, V2, V5) from the Chapman were used for training the VGG-16 Model as shown in figure4.2, to ensure overlap and parity between training data from the Chapman and MITBIH datasets.

Table 3.2: Merged class labels for both datasets [MITBIH & Chapman] and corresponding SNOMED CT codes

Chapman	**Merged Labels**	**MITBIH**	**SNOMED CT**
AFIB, AF	AFIB	AFIB, AFL	164889003, 164890007
SVT, AT, SAAWR, ST, AVNRT, AVRT	GSVT	VT, T, VFL, SVTA	427084000, 713422000, 233897008, 164895002, 111288001, 251180001, 426761007
SB	SB	SBR	426177001
SR, SI	SR	N	426783006
	Omitted	PREX, B, IVR, P, BII, NOD, AB	

3.2.1 Domain Data Pre-processing

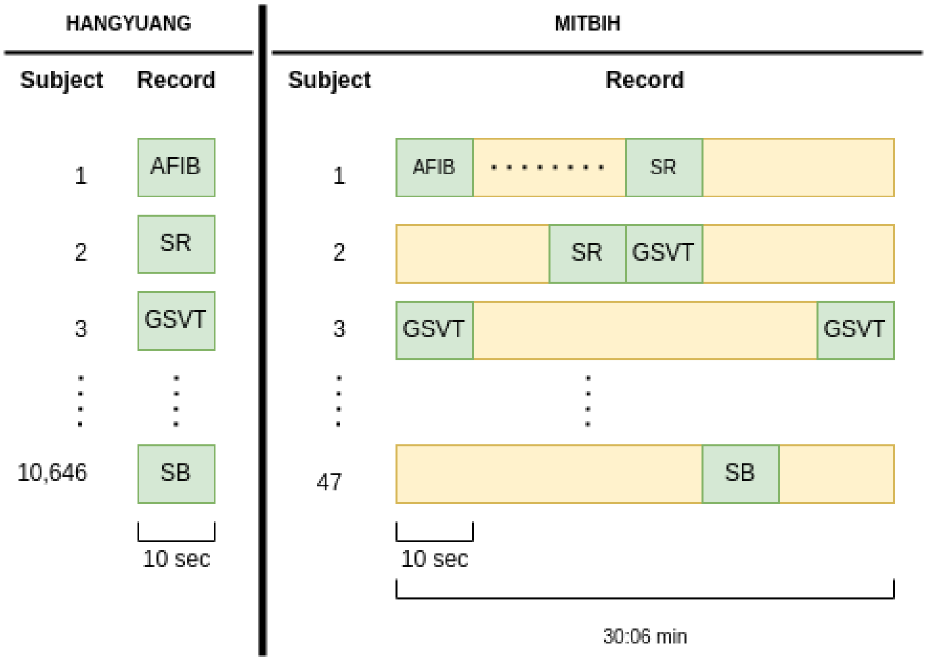


Figure 3.2: Data Windowing Visualization. In the Chapman [labelled, Hangyuang] dataset 10646 subjects contributed to 10646 10-second ECG records. In the MITBIH dataset 47 subjects contributed 47 30:06-minute records windowed to 7414 10-second ECG records.

The datasets shown in Figure 3.3 are shown with record labels already having been merged according to the Table 3.2.

Chapman Dataset:

The Chapman dataset was not processed beyond the relabelling of the records with the merged classes. The dataset was already pre-processed by the curators of the dataset. The same methodology used by the curators of the Chapman dataset to denoise and filter the signal was eventually implemented for the MITBIH dataset.

MITBIH Dataset:

The MITBIH dataset was originally 47 subjects with each subject having a single 30:06 min long record. Each 30:06 minute, 2 Lead ECG signal is truncated into a series of records of 10 second windows with no overlap between adjacent records. The individual signal annotations available for the MITBIH dataset were used to relabel each of the truncated windows with one of the 4 merged class labels. Some subjects would have had multiple different labels throughout the long signal and each truncated signal record was labelled with the preceding record's label unless there was a label change within the file itself. In case of multiple classifications in a single truncated record, the record was removed from consideration. Also, as shown in Table 3.2 there was a small list of classification labels that didn't fit in with any of the other 4 merged classes and were also omitted form consideration. After all the truncating and omissions 7414 total records were produced to be split between testing and training.

3.2.2 Signal filtering

As discussed in Section 3.2.1 the Chapman dataset signals were pre-processed by dataset curators [99] to produce a denoised signal, with all power noise, motion artifacts and random noise as thoroughly minimized as possible. The signal also had any DC signal offset (baseline wander) removed. This was done by implementing a series of very specific filters to accomplish this.

First [99] implemented a low pass butterworth filter with a passband from 0.5 Hz to 50 Hz and a stopband to 60 Hz with 2.5 db attenuation to remove power noise that exists at 60 Hz. This was followed by a LOESS (Local polynomial regression) smoother to remove the baseline wander of the signal. Lastly, a NLM (non local means) noise reduction algorithm was also used to smooth out the high frequency noise that might have appeared after the LOESS smoothing [99].

The denoising tools discussed are available in opensource matlab code available at `https://github.com/zheng120/ECGDenoisingTool`, provided by the curators of the dataset. All MITBIH 10 second windowed signals were passed through the denoising tool to achieve parity in denoising with the training dataset. Because the code in matlab was very slow to process and took a very long time to denoise even a single file with 2 ECG leads. The code was re-implemented in python using all the same parameters presented in great detail in the matlab tool.

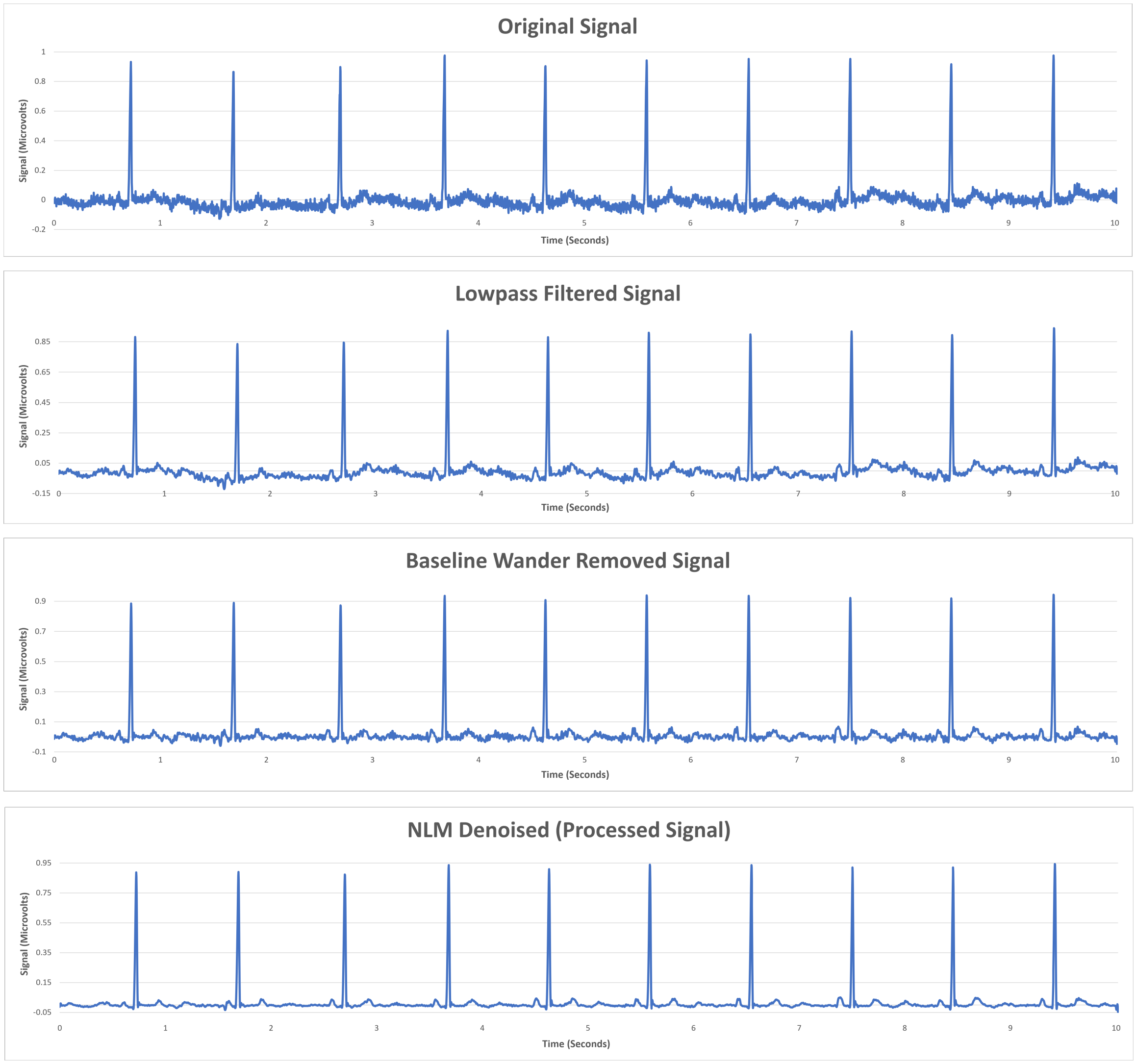
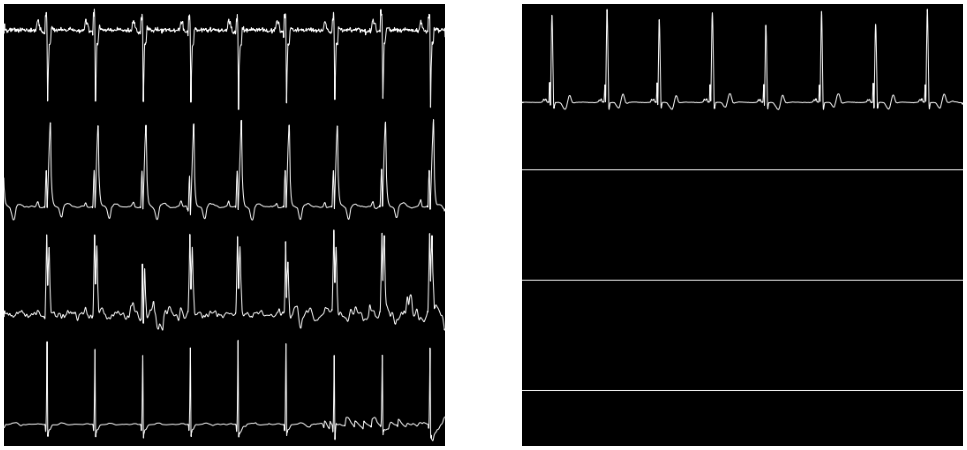


Figure 3.3: Signal filtering - Each step of the signal filtering process visualized. The Original Signal is passed through a Lowpass Butterworth filter, A LOESS filter to remove baseline DC wander, then NLM denoising to remove high frequency components introduced in the previous filtering efforts.

Once the signals were filtered and denoised, they were plotted as binary images with the background set to black or 0 and the signal's pixel values set to white or 1 of size 500 x 500 pixels. This image was plotted with only the ECG leads 'II, V1, V2, & V5' as these are the only combination of leads used by the MITBIH database (with the exception of MITBIH subject 124 that recorded has leads II % V4, and lead V4 was not plotted for consistency). As shown in figure 4.4

Following the plotting of the signals, the signals were once again filtered using threshold binary and OTSU filtering methods to remove any aliasing artifacts or noise from the plotting process. The reason the images were kept as plotted images and not just stored as an array from the beginning, was because the XAI methods discussed in Section 5.1 required images to be fed into some XAI methods and the plotted images were created for convenience of usage later in the process. The plotted image that had been filtered with threshold binary and OTSU filtering were stored in an sequential array, to be passed on to the CNN along with an array of matching labels. From this point onward, these will be referred to as 'plotted images'.



(a) Record from Chapman Dataset - Leads II, V1, V2, V5

(b) Record from MITBIH Dataset - Leads II, V4 and 2 empty leads

Figure 3.4: Potting Leads II, V1, V2, V5. These images show the visualization of records from each dataset [Chapman & MITBIH] presented to the AI model for training and testing.

Chapter 4

AI Model: ECG Classifier

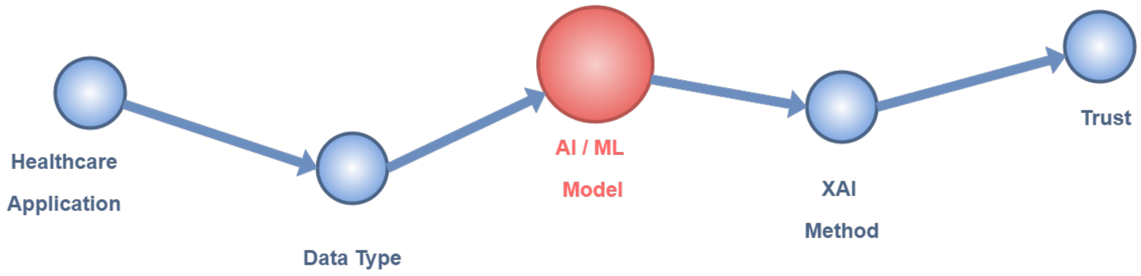


Figure 4.1: Visualization of overall Journey of the Thesis - Specifically dealing with the AI model in this section.

This section discusses how the underlying AI model is built, the rationale behind each component and the decisions for the parameters.

4.1 Model Architecture

The VGG-16 (a 16 layer, 2-D convolutional neural network) was used as the primary component of the AI model, along with a fully connected classifier connected at the final stage. [102, 103] Demonstrated the use of a variety of 2-D CNNs to perform classification tasks involving biomedical signals. [102] Utilized a two lead raw ECG signal and converted the signal into a 2-D binary image to be classified using the VGG-CNN architecture as a powerful feature detector. Similarly [103] demonstrated the use of 2-D VGG-CNN, among other CNN architectures, to be superior in performance compared to a traditional 1-D signal processing method commonly used with time-series data like ECG signals.

The benefit of using a deep learning architecture like VGG was the ability to utilize a well pre-trained neural network with many specialized, complex feature-detecting kernels. Greatly reducing the need to manually perform complex feature detection work prior to classification [104].

As shown in [102, 103] other similar architectures such as Inception, Xception, ResNet, AlexNet, could have been used as well. The VGG-16 model had a similar accuracy as the other models, but with a much larger number of trainable parameters. This made VGG-16 more expensive computationally relative to the ResNet, Xception and Inception models [105]. This model selection assumed that more learnable parameters, may allow for more detailed insights at each layer when the XAI methods are implemented. This assumption did not hold true, based on how XAI methods work. Other computationally lighter AI models could have been used instead.

4.2 Training Paradigm

The training paradigm was developed through a series of design iterations. The best combinations of model parameters (discussed in detail in the section 4.3) was determined with a combination of cross-validation and exploring literature. Graphical representation of all attempted training paradigms is presented in Appendix A as well as a full table outlining the results of the iterative testing (Appendix B, Table A.1). This sections presents a brief overview of the training paradigm of the CNN model.

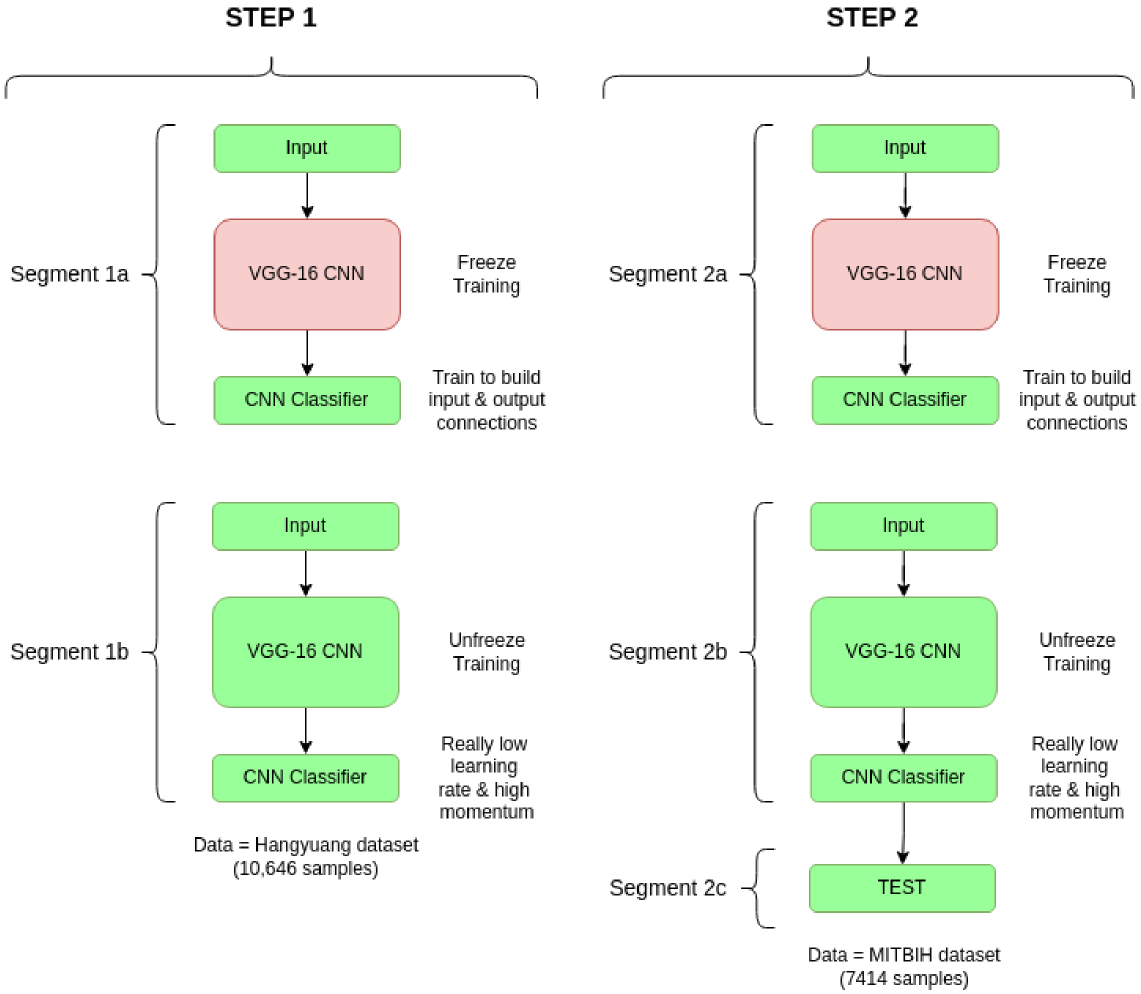


Figure 4.2: Overall training paradigm. Contains two steps of training, once with each dataset to achieve some degree of 'transfer learning' between datasets. Additionally each step involves initial training with frozen VGG-16 weights to initialize training of input and classifier layers, followed by unfrozen training for all layers of the whole model [with very low learning rates].

The training consists of 2 steps as shown in Section 4.2:

Step 1: Training with Chapman dataset.

Datasets characteristics are presented in the section 3.1.1. The machine learning model training on the Chapman dataset had 2 steps with separate segments:

1. In segment 1a, the parameters of all the layers of the VGG-16 CNN were kept 'frozen'. Meaning the weights of the neurons were untrainable. This was done to ensure that the input layer and the output dense (classifier) layers could be trained without changing the weights of the VGG-16 layers.

   The imagenet [97] pre-trained weights were used as the starting weights of the VGG-16 layer. As shown in [106] the imagenet pre-trained weights led to significantly improved performance in classification tasks. Additionally, given the relatively small size of the Chapman and MITBIH datasets, using imagenet pre-training utilized the feature detection power from the pre-trained convolutional filter kernels [107].

2. In segment 1b, all the parameters were unfrozen to allow slight changes in the weights of the neurons in the VGG-16 CNN layers. To prevent massive fluctuations in the weights, the learning rate was reduced [107] and the momentum was increased to counterbalance the impact of the low learning rate [72].

Step 2: Training with the MITBIH dataset.

The full process of **Step 1** was repeated with the MITBIH dataset. Details of the datasets are presented in section 3.1.1. Only 30% of the data from the MITBIH dataset is used for training and validation purposes and 70% is retained for testing. This second step was only done with the MITBIH dataset because this is the testing Dataset, the entire Chapman dataset was used for training only (as shown in Step 1).

4.3 Model Parameters

This section presents the hyperparameters introduced in section 2.2.3. In this section more detail about how they were implemented is presented.

Hyperparameters that were optimized included:

1. **Freezing/Unfreezing Training** - The training happens in 2 segments. First, all the convolutional layer parameters are set to be 'frozen' (untrainable), so that the learned weights don't change for these layers as the training progresses. Only the Input and final Dense (classifier) layers will have changing weights. This initial training is followed by the second training segment, where all layer parameters are unfrozen (trainable) to allow changes in the weights of all parameters.

2. **Loss function** - The 'categorical cross-entropy' loss function is used because the task is a multi-classification task and according to [108] & [109] categorical cross-entropy (log) loss function is the most commonly used loss function for a multi-class problem. Although more sophisticated loss functions exist [70], the simplicity of pre-existing keras implementation makes the log loss function the obvious choice .

3. **Optimizer** - The SGD (Stochastic Gradient Descent), this was the default optimizer implemented by Keras and recommended for a multi-classification problem [71].

4. **Learning Rate** - An initially high learning rate of 0.0001, to move the weights of the input layer and the dense (classifier) layer to the neighbourhood of the optimum weights quickly. In the second round the a much lower learning rate of 0.00001 is used to prevent massive fluctuations in trainable parameter weights.

5. **Momentum** - A low initial momentum of 0.5 for the first part (with frozen training layers), followed by an increased momentum of 0.9 when the learning rate

increases in the second cycle of training (with unfrozen training layers). The momentum is basically using an exponentially weighted average of the gradients used to determine the weights in each epoch, and utilizing this weighted gradient to determine the weights instead, which allows the neural network to arrive at the final weights quicker [72].

6. **Test, Train, Validation Split** - 100% of the Chapman data was used for training. 70% of the MITBIH dataset was set aside for testing. The remaining 30% of the overall dataset was used in the training stage and 10% out of the 30% training data was used for validation. Therefore 27% of the overall data was used for training and 3% for validation.

7. **Batch Size** - With larger batch sizes, greater computational resources are required and overall generalizability of the model decreases [110] but with smaller batch sizes

8. **Learning Epochs** - Based on preliminary testing, approximately 20 Epochs were determined to be long enough to see training and validation plateau and considered long enough to see if any further significant changes in gradient descent occur, resulting in further changes to the validation accuracy.

7. **Early Stopping Parameters** - The validation accuracy was the value that was monitored to implement early stopping. A minimum change or 'min_delta' of less than 1% in the value was the set point. As long as this set point didn't change more than the min_delta over a period or 'patience' of 5 consecutive epochs, the training was stopped. This method was implemented by the keras' 'EarlyStopping' method and is what helped determine the optimal cutoff of 20 epochs determined above [71].

Table 4.1: Model parameters used by the AI model for training purposes. These include parameters and hyperparameters that were optimized using values from cross validation and literature.

Parameter	**Segment**	**Value**	**Reference**
Freezing/Unfreezing	1a, 2a	Frozen	[107]
	1b, 2b	Unfrozen	
Loss Function	All	Gategorical Cross-Entropy	[108, 109]
Optimizer	All	Stochastic Gradient Descent	[71]
Learning Rate	1a, 2a	0.0001	[107]
	1b, 2b	0.00001	
Momentum	1a, 2a	0.5	[72, 111]
	1b, 2b	0.9	
Test, Train, Val. Split	1a, 1b	100%, 0%, 0%	Experimental
	2a, 2b	70%, 27%, 3%	
Batch Size	All	16	Experimental
Epochs	All	20 (with Early stopping)	Experimental
Early Stopping	All	Value - Validation Accuracy Min_delta - 1% Patience - 5	[71]

4.4 Training

The input images created from plotting the various lead combinations were of the size 500x500x1 pixels. The regular input of the VGG-16 is 224x224x3. To accommodate the difference size an input block was attached to the VGG-16 CNN layers to modify how the image features were extracted (see training paradigm in figure 4.2). New fully connected dense layers were added to process that increased number of parameters (as the input increased from 224x224x3 to 500x500x3). It also allowed the CNN to redefine the number of classes from 1000 (normally for VGG-16) down to 4 for ECG classification in this thesis. A full layer by layer output is presented in Appendix B Model B.

The input images were modified from a single layer to a three layer image (500x500x1 to 500x500x3) as required by the VGG-16. This was done by broadcasting the array of images to increase the dimensional space. When the initial training was done with frozen parameters, only the CNN parameters were kept frozen (14,714,688 parameters frozen out of 118,492,676 total parameters).

There was a transfer learning effect created by training the VGG-16 model on the Chapman dataset in step 1 (figure 4.2), and in step 2 training with 30% of the MITBIH dataset. The impact of this potential transfer learning effect was evaluated and discussed in model performance section 4.6. In addition to these considerations, class weighting was implemented when training with the MITBIH dataset to offset the large dataset imbalance [112].

An important consideration for training paradigms, which is commonly overlooked in many medical decision making AI models. It is understanding the intersection of subject (patient) & records (ECG recordings of a patient) (figure 3.3) and how they impact learning in AI models [113]. According to [113] when records from the same subject are intermixed between training, validation and testing stages they often result in a massive improvement of the prediction accuracy of the model. This is effectively cheating, because during the testing phase the model has likely seen records very similar to the test set from the same patient during training or validation. The best method for dealing with multiple records from the same subject is to isolate all the records for a given subject for either training, validation or testing stage alone.

This common oversight is very prevalent in literature with wearable sensors used to predict clinical outcomes.

Given the focus of this thesis' research was quantifying explainability in medical AI. The author acknowledges that no subject/record isolation was performed and that data leakage was not accounted when developing the model. As the goal of the thesis is to scrutinize model decision making at the the XAI explanation stage. An undeservedly-superior classifier may offer greater insight on how explanations are generated for an Neural Network's decision making paradigms.

To perform this task, a high accuracy model was beneficial to observe the best possible behaviour of the XAI methods. This particular consideration applied to the training done with the MITBIH dataset in step 2 of the training paradigm as shown in Figure 4.2. Having a model with high (artificially inflated) accuracy and a model with lower accuracy could have both been tested together to see their impact on the results of the XAI outputs, this could potentially have been used to gain insights on the stability of the XAI output with changing model accuracy. But this effect was not explored. If explainability is correlated to the accuracy of the model, as the accuracy changes, the similarity metric of explainability will likely change proportionally, but the within XAI method stability should most likely remain the same, and this may give additional insights on the value of the XAI methods.

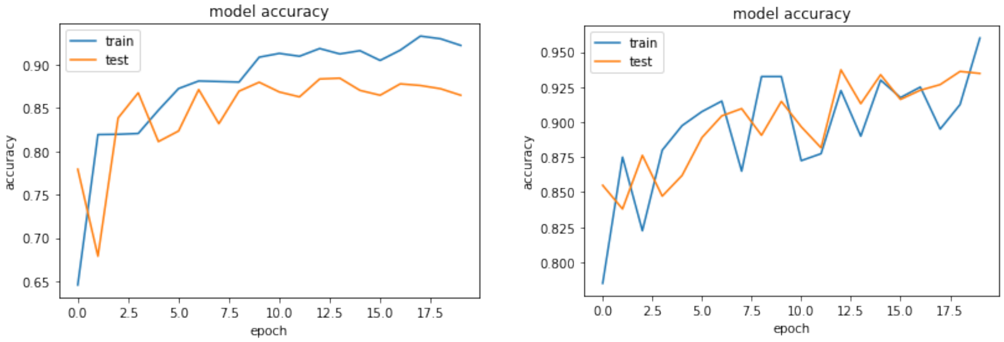
4.5 Testing

The only dataset used for testing is the MITBIH dataset, 70% of the dataset (5190 records) was allocated for testing. As discussed before, the subjects seen in the training phase might be seen again in the testing phase, but the individual records would be unique and never before seen.

Just as in the training, testing the performance of VGG-16 based CNN was done with the plotted images. A list of these files seen here was saved, to be given to the XAI methods in section 5.1.

4.6 Model Performance

Step 1 The results recorded here are of the performance of the model with the Chapman dataset only.

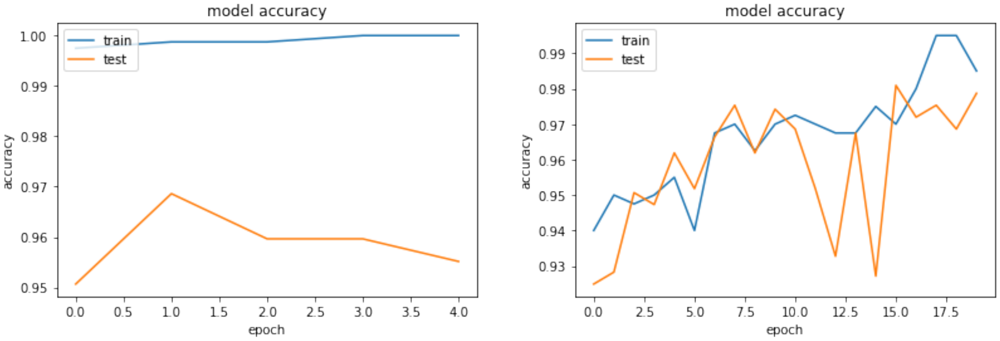


(a) Training Step 1, Segment 1a Performance (Accuracy vs. Epochs)

(b) Training Step 1, Segment 1b Performance (Accuracy vs. Epochs)

Figure 4.3: Training Step 1 - Performance (Accuracy vs. Epochs)

Step 2 The results recorded here are of the performance of the model with the MITHBIH dataset, after the model has already been previously trained on the Chapman dataset.



(a) Training Step 2, Segment 2a Performance (Accuracy vs. Epochs)

(b) Training Step 2, Segment 2b Performance (Accuracy vs. Epochs)

Figure 4.4: Training Step 1 - Performance (Accuracy vs. Epochs)

Results

Table 4.2 shows the accuracy, precision, recall and Area under the curve (AUC) values of the model during Training, Validation [step 2, segment 2b] and Testing. More detailed values shown in the Appendix A table A.1.

Table 4.2: Model Performance Results

Train				**Validation**				**Test**			
Train Acc.	**Precision**	**recall**	**auc**	**Val Acc.**	**Precision**	**recall**	**auc**	**Test Acc.**	**Precision**	**recall**	**auc**
0.985	0.97	0.97	0.9904	0.9787	0.9636	0.9507	0.9966	0.9718	0.9475	0.9391	0.9943

Confusion Matrix

The confusion matrix shown in table4.3 displays results of the Testing with 5190 records. Each quadrant with a Class (AFIB, SB, GSVT, SR) represents the results for that classification. In the rows in each quadrant, '1' represent an expected result of that classification and '0' represents an expected result of any classification other than that quadrant. The column '1' represents a correct classification and '0' represents an incorrect classification event.

Table 4.3: Confusion matrix for test results, each class (AFIB, SB, GSVT, SR) is presented separately in a quadrant

AFIB	**0**	**1**	**SB**	**0**	**1**
0	4561	46	**0**	5065	3
1	11	572	**1**	4	118
GSVT	**0**	**1**	**SR**	**0**	**1**
0	4869	115	**0**	785	126
1	121	85	**1**	154	4125

4.7 Software Packages

The full flowchart of methodology and each .ipynb file in it is provided in Appendix B Figure A.6.

Command to get all the requirements.txt files for all the jupyter notebooks used:

```
jupyter nbconvert --output-dir="./reqs" --to script filename.ipynb
cd reqs
pipreqs
```

1. matplotlib==3.1.2
2. neurokit2==0.1.1
3. numpy==1.17.4
4. opencv_contrib_python==4.5.3.56
5. pandas==0.25.3
6. tensorflow==2.7.0
7. tensorflow_gpu==2.6.0
8. wfdb==3.3.0

Chapter 5

XAI Methodology

5.1 XAI Analysis

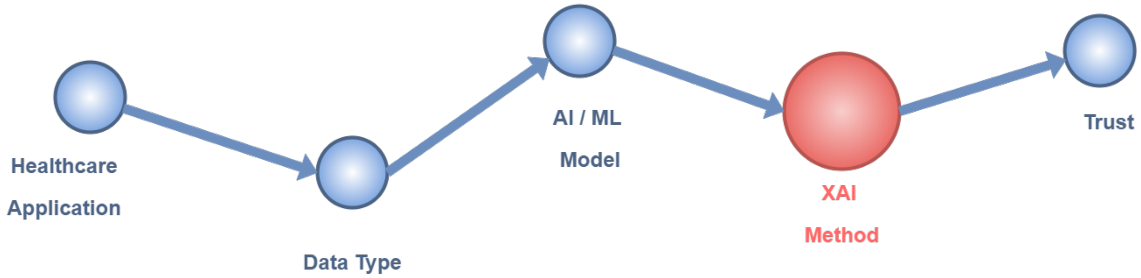


Figure 5.1: Visualization of overall Journey of the Thesis - Specifically dealing with the XAI methods in this section.

This section presents the implementation of specific XAI methods and the techniques used to provide: 1) a comparison between XAI methods, and 2) a comparison between XAI vs. SME.

5.1.1 Pseudocode of XAI output processing pipeline

The following steps were taken to perform the XAI analysis and the subsequent subsections expand on each step:

Step 1: MITBIH processed image output directory was sorted to identify testing images.

Step 2: XAI outputs were generated for each test image.

Step 3: XAI outputs were processed (thresholded, filtered, & masked).

Step 4: Properties for clusters of interest (centroids, & boundaries) were identified

Step 5: Comparisons of the processed XAI outputs were performed.

5.1.2 Step 1 - Sorting and Identification of MITBIH images

A directory of all MITBIH images used by the ECG classifier model were sorted to identify images used for testing, because they were originally randomly distributed using the test_train_split function. A list of indices used for testing were identified and only those plotted images were passed to the XAI method in step 2. Algorithm 1 outlines the logic discussed in this subsection.

```
Algorithm 1: Tracking indices of sorted files
Input  : Directory of all MITBIH plotted images
Output: XAI input images
$TestTrainSplitfunction(input) \leftarrow list\_of\_test\_file\_indices$
$Sort(list\_of\_test\_file\_indices)$
for $input$ do
    if input is in list_of_test_file_indices then
        $Load(input) \leftarrow images$
        $Preprocess\_input(images) \leftarrow XAI\_input\_images$
        $XAI\_input\_images \leftarrow$ pass to XAI method
    end
end
```

5.1.3 Step 2 - Generating XAI outputs

From step 1 an array of XAI input images was passed to one of four XAI methods: Vanilla Saliency, SmoothGrad, GradCAM, & GradCAM++. XAI methods were obtained from tf-Keras-vis [1] [114]. Attention maps were produced for each input image provided to the XAI methods. Used to visualize ares of high interest to the ECG classifier model (VGG-16) in its decision making process.

Algorithm 2 references the XAI methods: Vanilla Saliency, SmoothGrad, GradCAM, & GradCAM++, these methods are discussed in more detail in section 2.3. The activation function of the last CNN layer was altered to be a linear function. The output became directly proportional to the input only modified by the weights of hidden layer neurons, and not a complex mapping of a non-linear function. This

[1] `https://keisen.github.io/tf-keras-vis-docs/`

provided a linear combination of all the neuronal weights and biases represented by a Taylor series. Magnitudes of the weights relative to pixels of the original image created a mapping of importance of the pixels in determining the output score [114].

Algorithm 2: Generating XAI outputs

```
Input  : XAI input images
Output: XAI method specific visualization of attention map
$ReplaceToLinear$ modifies last Conv. layer activation function to
  linear
$Score\_function \leftarrow score$ (links output of last Conv. layer to
  predicted classification)
for $input$ do
    $Vanilla\_Saliency(input, score) \leftarrow vs\_XAI\_output$
    $SmoothGrad(input, score) \leftarrow sg\_XAI\_output$
    $GradCAM(input, score) \leftarrow gc\_XAI\_output$
    $GradCAM++(input, score) \leftarrow gc\_XAI\_output$
end
$XAI\_outputs$ passed to post-processing
```

Figure 5.2 shows attention map outputs generated by each XAI method listed in Algorithm 2. These outputs were generated for the same arbitrary sample input.

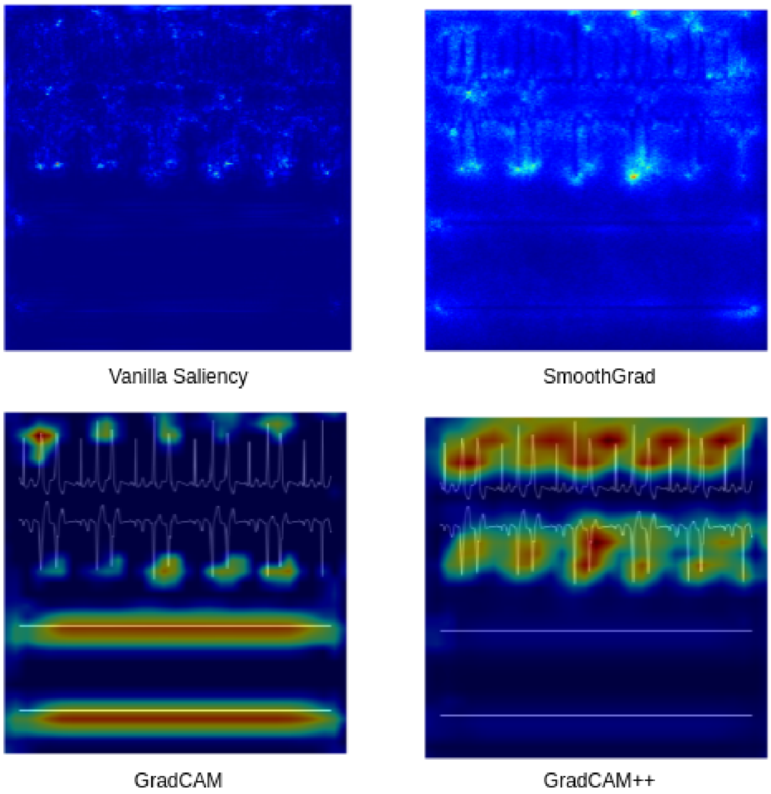


Figure 5.2: Visualization of steps from Algorithm 2 - XAI generated attention map outputs from 4 different XAI methods [Vanilla Saliency, SmoothGrad, GradCAM & GradCAM++

5.1.4 Step 3 - Processing XAI outputs

The XAI output images generated in Algorithm 2 were passed to the next stage for processing before segmenting the image. The images were thresholded into binary images using Otsu thresholding (see Appendix A Figure A.6) to ensure uniform processing of the image regardless of the XAI method colour map.

To filter the image, first, a median filter was applied to reduce noise. Unlike low-pass filtering, commonly used to reduce grain noise, median filtering allowed smoothing over an area with drastic differences in pixel values [115]. Median filtering was done in two discrete stages with increasing footprint sizes to allow fine denoising (small filter footprint) followed by coarse denoising (larger filter footprint) [115].

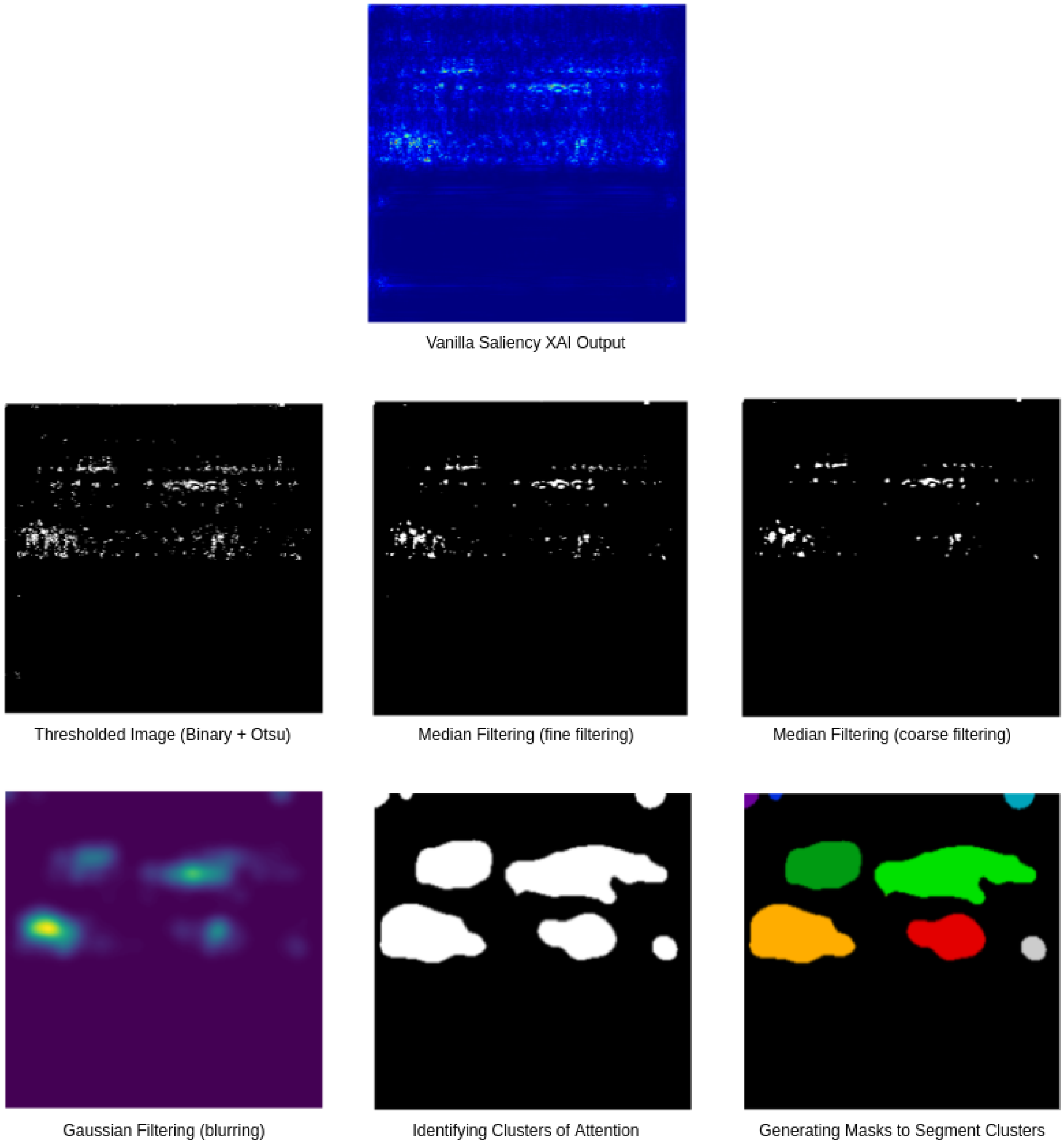
This was followed by Gaussian filtering, used to blur (smoothen) the edges of the clusters of interest in the image. The gaussian filter kernel applied a heavier weighting to the central region of the cluster relative to the periphery, and smoothed its edges [116]. Multi-stage filtering was preferred over single stage filtering. This removed high frequency noise and reduced pepper noise, commonly observed in XAI outputs. It also allowed a fine control over preventing the segments of interest from growing too large [115] [116]. After filtering the remaining clusters were segmented. These segmented clusters were passed to step 4.

The Gaussian filter allowed blurring and consolidating clusters of interest close to one another but distant beyond a few pixels. Gaussian filtering was used to reduce the total number of clusters to be analyzed. In preliminary testing it was noted that the XAI outputs often appeared similar to the SME annotations, except they were scattered in many small clusters. These neighbouring clusters were consolidated using the Gaussian blur filter.

Algorithm 3: Processing XAI outputs

```
Input  : XAI method specific visualization of attention map
Output: Filtered and Segmented XAI output
$Thresholding(Binary + Otsu)$ (thresholds the XAI attention map to unify filtering regardless of XAI method)
$Median\_filtering\_1 \leftarrow fine\_denoising$ (smoothes and consolidates regions of grainy noise)
$Median\_filtering\_2 \leftarrow coarse\_denoising$ (separates distant larger sections and removes excess noise)
$Gaussian\_filtering \leftarrow smoothing\_regions$ (groups nearby neighbouring sections)
$Masking \leftarrow segmented\_XAI\_output$ (creates masks on each distinct cluster)
for $input$ do
    $XAI\_segmentation\_function(input) \leftarrow Filtered\_and\_Segmented\_XAI\_output$
end
$XAI\_outputs$ passed to identify contour data
```

Figure 5.3 below shows a visual representation of the outputs generated at each step of Algorithm 3. In the figure, a single vanilla saliency XAI output is chosen as an arbitrary example.



Vanilla Saliency XAI Output

Thresholded Image (Binary + Otsu)

Median Filtering (fine filtering)

Median Filtering (coarse filtering)

Gaussian Filtering (blurring)

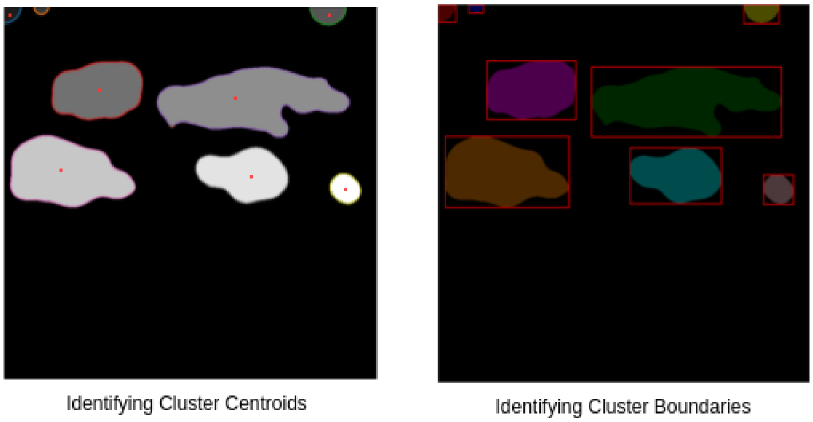
Identifying Clusters of Attention

Generating Masks to Segment Clusters

Figure 5.3: Visualization of steps from Algorithm 3 - Sequence of thresholding, filtering and segmenting steps to generate clusters of attention from an XAI output (Vanilla Saliency shown here).

5.1.5 Step 4 - Identifying Properties of Interest

The centroid and boundary positions were determined for each segmented cluster. The X and Y coordinates of the centroid, as well as the coordinates of the left, right, top and bottom most extreme positions of each segmented cluster were taken relative to the origin position of the image (top left corner). The coordinates were stored in a data-frame, and the areas of attention determined by XAI outputs were compared against the SME determined areas of attention in the next step. Figure 5.4 shows a visual representation of centroids and boundary positions around clusters of attention. Any clusters with Y coordinates corresponding to a location on the ECG record with no ECG lead data (II, V1, V2, and V5) were ignored as noise from the XAI output method.



Identifying Cluster Centroids

Identifying Cluster Boundaries

Figure 5.4: Visualization of centroid and boundary positions of each segmented cluster identified in figure 5.3

5.1.6 Step 5 - Comparison of XAI Methods

After the properties of interest for each image output by each XAI method are identified. The XAI methods were evaluated in an objective & rigorous manner to quantify trustworthiness. Section 5.2 presents details of the comparisons of XAI methods.

5.2 Comparison of XAI Methods

As established earlier in Section 2.3.3, trustworthiness of XAI generated explanations is broken into constituent components of **similarity** & **stability** and if the XAI can identify when there is **novelty** and what was **overlooked** in what the AI model has learned, all relative to the golden standard, subject matter experts.

5.2.1 Similarity

To test similarity between XAI outputs and SME annotations for a given record, SME annotation timestamps were converted to X-axis pixel positions for each corresponding ECG record (500x500 pixel). This process was performed for all rhythm and beat annotations by the SME, and the annotations positions were recorded in an array.

An XAI identified cluster from step 5.1.5 and SME annotation were considered a positive match if the SME annotation was anywhere between the left and right boundaries of the cluster. All individual clusters were compared against all SME annotations for each record. An array of pairwise values of cluster centroids & matching SME annotations was generated. Clusters with no matching SME annotations were recorded with a value of 0 for the SME annotation and vice versa. Two distance metrics, Jaccard and Hamming, were applied to each pairwise array to provide a measure of similarity between the XAI output and the SME annotations for each record. These distance metrics were applied to outputs from all XAI methods (Vanilla Saliency, SmoothGrad, GradCAM, and GradCAM++) separately for each record.

Equation 5.2.1 shows a visualization of the XAI output array (centroid ± boundary), SME annotations and resultant pairwise array (XAI_output, SME_annotation).

$$
\begin{aligned}
&XAI\_output = \begin{bmatrix} 5 \pm 2 & 26 \pm 5 & 151 \pm 6 & 225 \pm 3 & 350 \pm 10 \end{bmatrix}\\
&SME\_annotations = \begin{bmatrix} 1 & 157 & 222 & 400 \end{bmatrix}\\
&pairwise\_array = \begin{bmatrix} (0,1) & (5,0) & (26,0) & (151,157) & (225,222) & (350,0) & (0,400) \end{bmatrix}
\end{aligned}
\tag{5.2.1}
$$

5.2.2 Stability

To test stability of an XAI output, the performance of each XAI output was compared with all other XAI outputs for each individual record, in a pairwise comparison. An array of clusters from one XAI method was compared to an array from a second XAI method. Each cluster from the first XAI method was compared against all other clusters from the second XAI method. A positive match was considered if the centroid x-axis position for a cluster was anywhere between the left and right boundaries of the cluster from the second array. An array of pairwise values of matching cluster centroids was generated. Clusters from one array with no matching cluster from the second array were recorded with a value of 0 for the second array. and vice versa. The Pearson correlation distance metric was applied to each pairwise array to provide a measure of stability between the pairs of XAI outputs.

Equation 5.2.2 shows a visualization of the XAI output arrays (centroid ± boundary) from 2 XAI methods and resultant pairwise array (XAI_output, SME_annotation) for a single ECG record.

$$\begin{aligned}
XAI\_output\_1 &= \begin{bmatrix} 5 \pm 2 & 26 \pm 5 & 151 \pm 6 & 225 \pm 3 & 350 \pm 10 \end{bmatrix} \\
XAI\_output\_2 &= \begin{bmatrix} 10 \pm 7 & 30 \pm 1 & 151 \pm 100 & 400 \pm 10 \end{bmatrix} \\
pairwise\_array &= \begin{bmatrix} (5, 10) & (26, 30) & (151, 151) & (225, 151) & (350, 0) & (0, 400) \end{bmatrix}
\end{aligned} \tag{5.2.2}$$

5.2.3 Novelty

XAI methods were used to identify novel and overlooked learning by the underlying AI; this provided insights about the AI's performance. These insights were acquired by evaluating individual features uniquely highlighted by or overlooked by the XAI output relative to the SME annotations.

Figure 5.5 demonstrates how all possible novel and overlooked features are determined. Overlapping features between the array of XAI outputs and SME annotations were removed. Two resultant arrays were generated; a novel features array and an overlooked features array. The novel features array was made from features unique to the XAI output and the overlooked features array was made from features unique to SME annotation array.

The TF-IDF method (see Section 2.3.3) was applied to the new resultant arrays for every record. A TF-IDF score was calculated for each individual feature for each record. If available, the most likely novel feature and the most important overlooked feature from each record were passed on to the SME for further validation.

A low TF-IDF score meant that feature occurred more commonly in the entire collection of records (all ECG records for a given classification). Thus a feature in the novel feature array with low TF-IDF score was very likely to be a real novel feature rather than noise because it appeared repeatedly in many ECG records for the same classification. If a record had multiple novel features, the feature with the lowest TF-IDF score was identified as the most likely novel feature.

A high TF-IDF score for a feature meant that it was a unique identifying feature for that particular record. Features with high TF-IDF scores identified the most important overlooked features, from the remaining features not yet learned by the AI. If a record had multiple overlooked features, the feature with the highest TF-IDF value in the record was identified as the most important overlooked feature.

In figure 5.5, visualization of inputs shows an ECG signal (labelled Expert Annotations). The SME annotations are marked on the ECG signal and represented below the image in an array. The image to the right (labelled XAI Outputs) shows the locations of all XAI identified clusters of interest. The array below that image is a representation of all XAI output clusters of interest.

Figure 5.5 then shows the two representative arrays are being subtracted from each other to remove any values common between the XAI output and the SME annotation. The two resultant arrays are then passed to the TF-IDF method.

Finally, visualizations of outputs shows an example of novel and overlooked features on overlays of the XAI output + ECG signal. Overlooked Features figure shows a green line to mark overlooked feature in the ECG signal that was never learned by the AI. Using TF-IDF it is identified as the most important overlooked feature for that record. Novel features figure shows a red line highlighting a novel feature identified by the XAI. It has the lowest TF-IDF score and most likely identifies a novel feature learned by the AI.

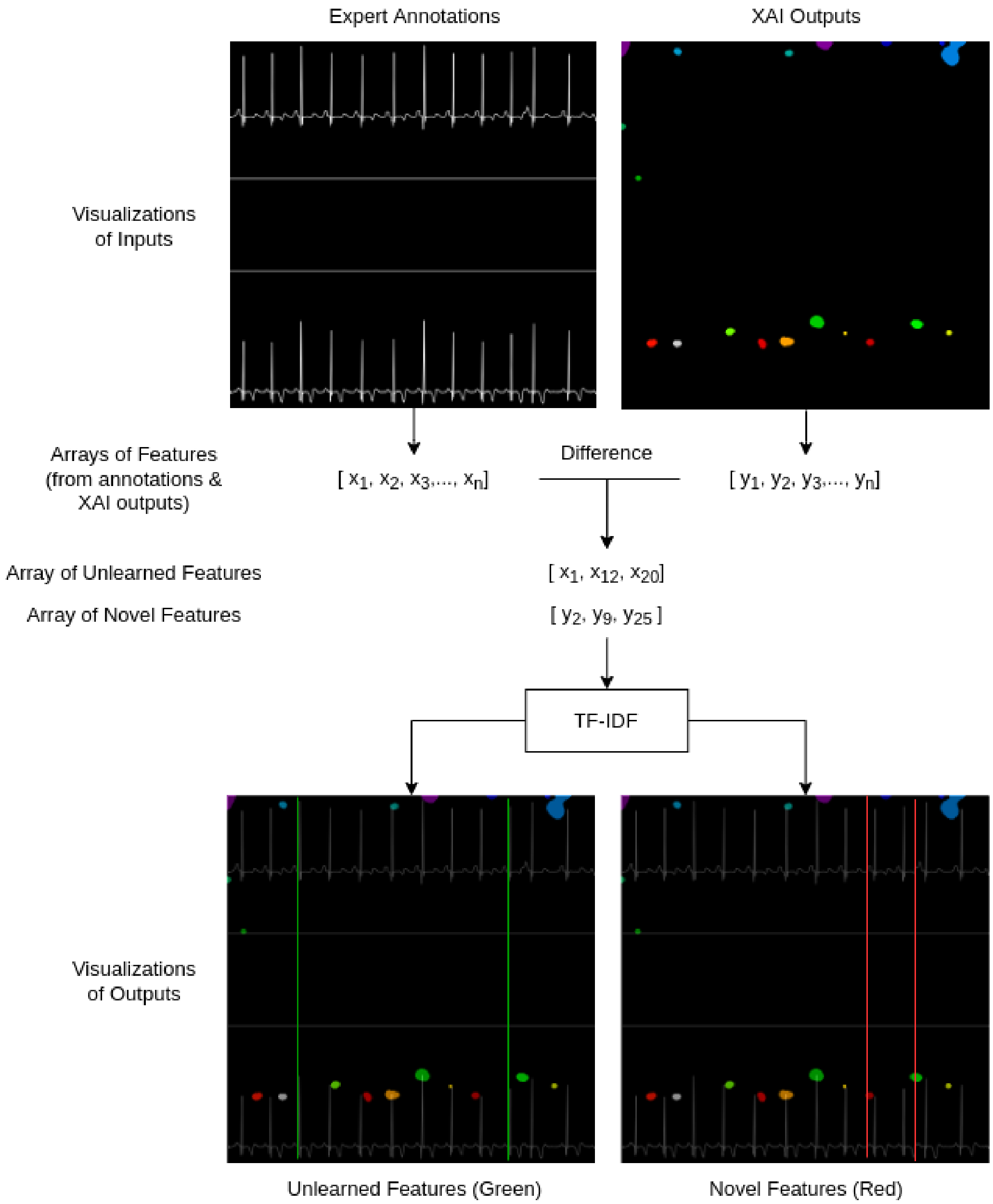

Expert Annotations
XAI Outputs
Visualizations of Inputs
Arrays of Features (from annotations & XAI outputs)
[$x_1, x_2, x_3, ..., x_n$]
Difference
[$y_1, y_2, y_3, ..., y_n$]
Array of Unlearned Features
[x_1, x_{12}, x_{20}]
Array of Novel Features
[y_2, y_9, y_{25}]
TF-IDF
Visualizations of Outputs
Unlearned Features (Green)
Novel Features (Red)

Figure 5.5: Visualization of how novel and overlooked features are determined. Arrays of features are subtracted from one another to remove all common features, remaining items in each array are evaluated using TF-IDF (shown in Figure 2.9) used to identify the significance of the possible novel and all overlooked features.

5.3 XAI Performance Results

This section presents the results of the comparisons of XAI methods with other XAI methods and SME annotations discussed in section 5.2. The results of evaluating XAI methods for similarity, stability and novelty are presented in the following section. Presented here are overviews and summaries of the results with more detailed tables and complete results available in the appendix A.

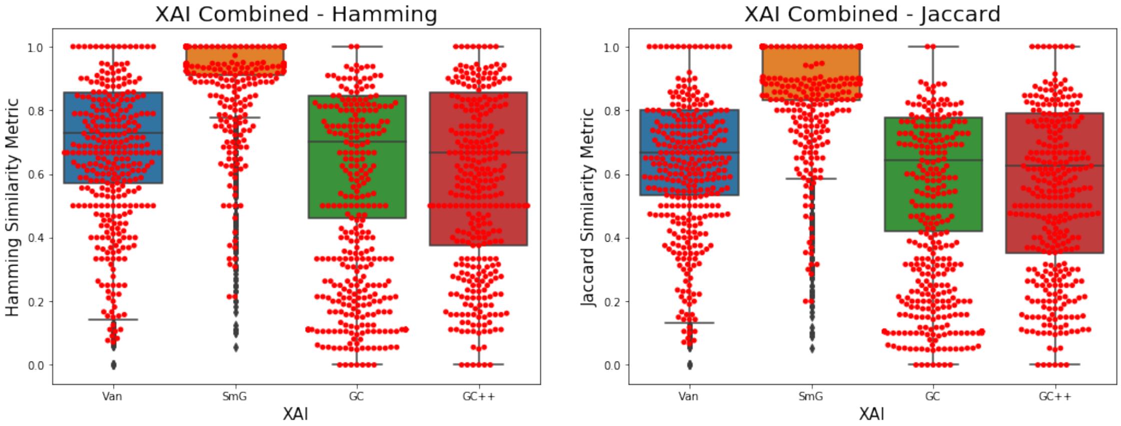
5.3.1 Similarity

The similarity XAI comparison method from section 5.2.1 compared the XAI output and the SME annotation to objectively quantify how accurately each XAI method captured and presented the underlying learning done by the AI model.

Table 5.1: Measuring Similarity Between XAI Output and Expert Annotation

XAI	Metric	Correct	AFIB_μ	GSVT_μ	SB_μ	SR_μ	Average	Combined Avg.
Vanilla Saliency	**Hamming**	**Y**	0.67862	0.55041	0.61465	0.70683	0.69880±0.20444	0.69470±0.20608
		N	0.69230	0.65855	0.84617	0.59571	0.63031±0.22098	
	Jaccard	**Y**	0.63994	0.51625	0.56646	0.65951	0.65267±0.19852	0.64916±0.19990
		N	0.65270	0.61900	0.80953	0.56263	0.59411±0.21301	
Smooth Grad	**Hamming**	**Y**	0.89149	0.92248	0.86749	0.92899	0.92306±0.13738	**0.92105**±0.13903
		N	0.90207	0.91033	1.00000	0.86922	0.88949±0.15964	
	Jaccard	**Y**	0.86558	0.89688	0.82027	0.90367	0.89714±0.15448	**0.89502**±0.15570
		N	0.87376	0.88262	1.00000	0.84084	0.86166±0.17050	
GradCAM	**Hamming**	**Y**	0.53191	0.56970	0.58998	0.65566	0.63842±0.24863	0.62719±0.25524
		N	0.50086	0.54146	0.60897	0.36973	0.45102±0.29102	
	Jaccard	**Y**	0.49715	0.52805	0.53929	0.60721	0.59160±0.23444	0.58125±0.24022
		N	0.46593	0.50082	0.53333	0.34553	0.41889±0.27016	
GradCAM++	**Hamming**	**Y**	0.50235	0.52619	0.64596	0.62432	0.60921±0.28916	**0.60477**±0.28853
		N	0.57985	0.61452	0.45513	0.46767	0.53509±0.26967	
	Jaccard	**Y**	0.47083	0.49021	0.59750	0.58101	0.56725±0.27473	**0.56321**±0.27399
		N	0.53875	0.57318	0.39047	0.43868	0.49990±0.25431	

Table 5.1 presents a summary of the outputs of each XAI method (Vanilla Saliency, SmoothGrad, GradCAM & GradCAM++) evaluated using the two metrics (Hamming and Jaccard) presented in Section 5.2.1. The 'correct' column separates test data by correct or incorrect classification decisions made by the underlying AI model. On the table 5.1 the highest combined average scores are highlighted in green, and the lowest scores are highlighted in red.



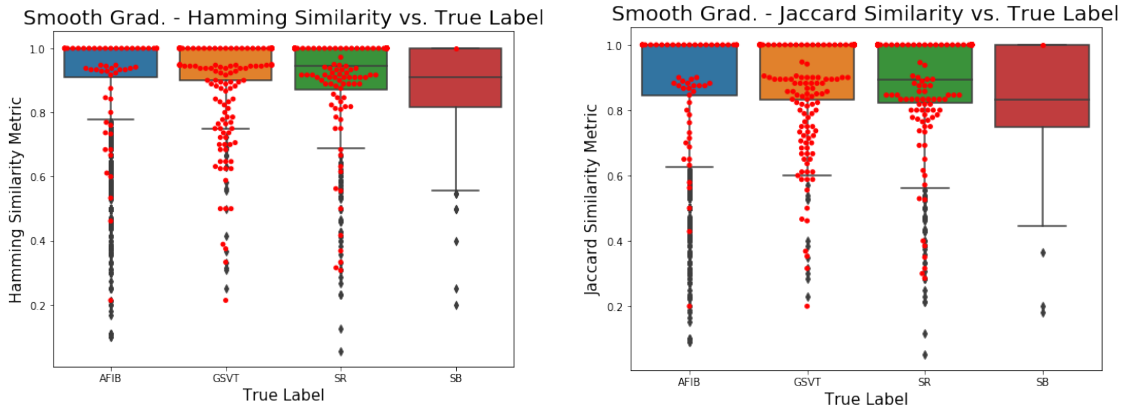
(a) Box-plot of all XAI outputs Hamming similarity scores for test ECG records.

(b) Box-plot of all XAI outputs Jaccard similarity scores for test ECG records.

Figure 5.6: Distribution of Similarity metrics of test ECG records

Figure 5.6 shows the performance of XAI methods combined over all classifications for each of the two similarity metrics. The BW plot shows the distribution of all test records and the red dots overlay the positions of only the incorrectly classified records.

The SmoothGrad XAI method had the highest Hamming and Jaccard combined average score across all classes (AFIB, GSVT, SB, SR), GradCAM++ has the lowest combined averaged scores. SmoothGrad also had the smallest standard deviations for each metric relative to all other XAI methods, and GradCAM++ had the largest standard deviations for each metric. Figures 5.7 and 5.8 visualize the scores of SmoothGrad and GradCAM++ in more detail via a BW plot.

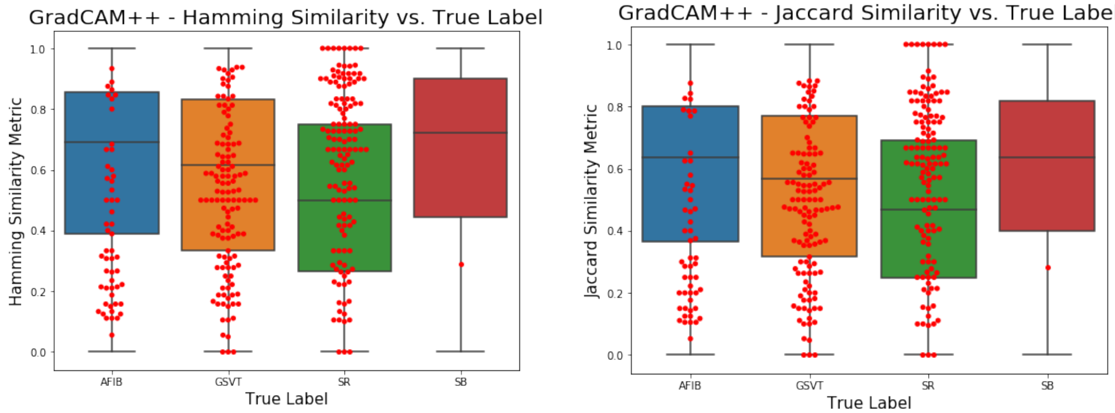


(a) Box-plot of XAI output vs.SME annotation compared with Hamming similarity metric for test ECG records.

(b) Box-plot of XAI output vs.SME annotation compared with Jaccard similarity metric for test ECG records.

Figure 5.7: Distribution of SmoothGrad similarity metrics of test ECG records vs. Classification labels

Figures 5.7a and 5.7b are the BW plots for Hamming and Jaccard similarity metrics respectively applied to Smooth Grad XAI method outputs. The upper whisker, 75th and the median values are near 1.0 for AFIB and GSVT. Exact values for the BW plots are available in a full detailed table in the appendix in Table A.3.



(a) Box-plot showing distribution of Hamming Similarity metric between the XAI output and the Expert annotation for each individual patient record in the test set.

(b) Box-plot showing distribution of Jaccard Similarity metric between the XAI output and the Expert annotation for each individual patient record in the test set.

Figure 5.8: GradCAM++ XAI output similarity metrics for each record plotted against True labels

Figures 5.8a and 5.8b are the BW plot for Hamming and Jaccard similarity metric respectively applied to GradCAM++ XAI method. The IQR range between similarity scores for GradCAM++ was large, similar to the large standard deviation seen in table 5.1.

The BW plots presented a comparison of accuracy of the two XAI methods (SmoothGrad and GradCAM++) by having the XAI methods output what they identified as the relevant learned features by the underlying AI and comparing that output with SME annotations. In figure 5.7 incorrect classifications by the AI (red points) were mostly below the median value and clustered around the 25th percentile or below in each BW plot. Whereas in figure 5.8 the incorrect classifications by the underlying AI model were evenly distributed throughout the BW plot. Many incorrect classifications had high similarity scores on Hamming and Jaccard metrics. The order of XAI methods from greatest similarity to the SME annotations to least similar:

Smooth Grad >Vanilla Saliency >GradCAM >**GradCAM++**.

5.3.2 Stability

To test stability of the four XAI methods (1-Vanilla Saliency, 2-SmoothGrad, 3-GradCAM, and 4-GradCAM++), pairwise comparisons between all XAI methods were performed for each individual record using pearson correlation (see section 5.2.2). The pearson correlation score (PCr) of each XAI method pairwise comparison was recorded in the table 5.2. For example PCr_1v2 refers to the pearson correlation coefficient between XAI method 1 (Vanilla Saliency) and XAI method 2 (Smooth Grad).

Table 5.2: Stability measure between XAI outputs, using Pearson Correlation

Record	Filename	y_True	y_Pred	Correct	PCr_1v2	PCr_1v3	PCr_1v4	PCr_2v3	PCr_2v4	PCr_3v4
0	file100x000	SR	SR	Y	1.00000	1.00000	1.00000	1.00000	0.68976	1.00000
1	file100x001	SR	SR	Y	1.00000	0.99999	0.59985	1.00000	0.81734	0.36977
2	file100x002	SR	SR	Y	1.00000	0.99999	0.51003	1.00000	0.44811	0.39226
3	file100x004	SR	SR	Y	1.00000	0.88922	0.80477	0.81766	0.82795	0.99777
...	...	...	...	...	...	...	...	...	...	...
5185	file234x173	SR	SR	Y	1.00000	0.81143	0.77011	0.64923	0.74210	0.57380
5186	file234x175	SR	SR	Y	1.00000	0.99981	0.52400	0.99846	0.56143	0.99854
5187	file234x177	SR	SR	Y	1.00000	0.90303	0.66749	0.83292	0.75267	0.00000
5188	file234x178	SR	SR	Y	1.00000	0.99973	0.99973	0.78991	0.99999	1.00000
5189	file234x180	SR	SR	Y	1.00000	0.75627	0.99998	0.63928	0.86787	1.00000
Average					**0.96660**	**0.81381**	**0.81031**	**0.79948**	**0.78724**	**0.77011**
Std. Dev					**0.09405**	**0.17844**	**0.20372**	**0.20174**	**0.23406**	**0.30996**

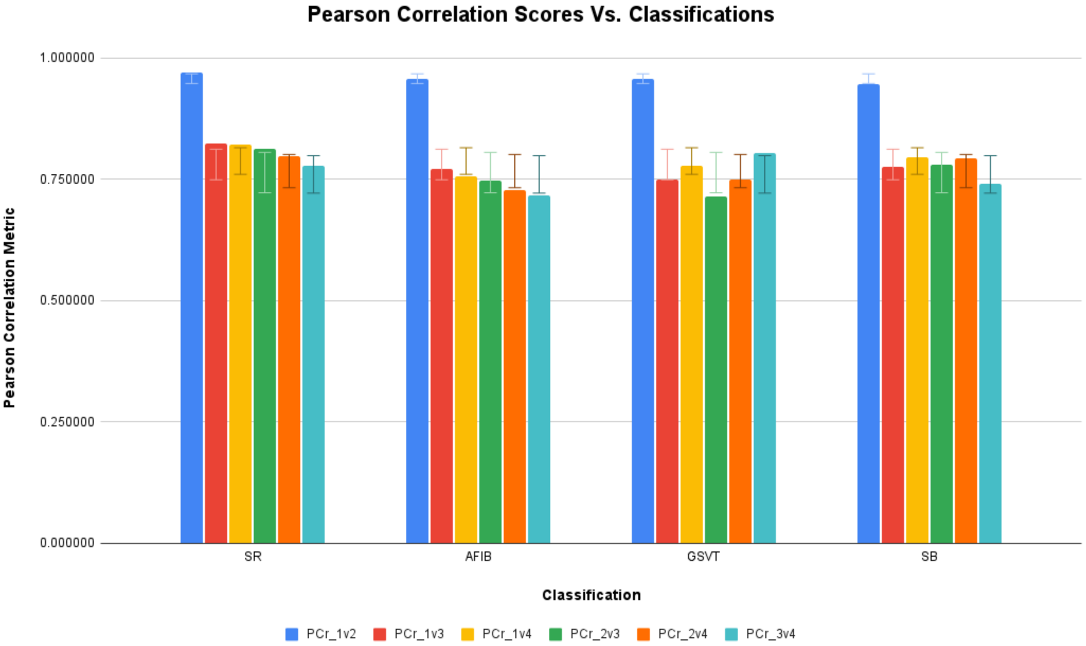


Figure 5.9: PCr Scores of Each XAI Method Pairwise Comparison vs. Test Record Classification

After pairwise comparisons were made between the XAI outputs of individual test records, the PCr scores were averaged. For example, to determine the relative stability of XAI method 1 (Vanilla Saliency) compared to XAI method 2 (SmoothGrad), the individual scores of PCr_1v2 were averaged (0.96660 ± 0.0945). Table 5.2 is a truncated table that shows the salient components used to determine the stability scores.

The highest PCr (average over all test records) between two XAI methods was 0.96660 for PCr_1v2 (between XAI methods 1 & 2, Vanilla Saliency and SmoothGrad respectively), highlighted in green in table 5.2. The lowest individual pairwise stability score was 0.77011 for PCr_3v4 (between XAI methods 3 & 4, GradCAM and GradCAM++ respectively), highlighted in red.

Figure 5.9 is a visualization of PCr scores shown in table 5.2 separated by record classes (y_True in table 5.2). The PCr between XAI 1 & 2 was significantly higher compared to all other pairwise comparisons. PCr_1v2 also had a small standard deviation, meaning far less variance in the XAI outputs of the Vanilla Saliency and SmoothGrad methods relative to all other pairwise comparisons.

Table 5.3: Overall Stability Scores for Each XAI method

XAI	Avg. Pearson Corr. Coeff.	St. Dev
Vanilla Saliency	**0.86357**	±0.18084
Smooth Grad	0.85111	±0.20363
GradCAM	0.79447	±0.23775
GradCAM++	**0.78922**	±0.25374

The PCr scores for a single XAI method for all pairwise comparisons were combined and averaged. For example, all scores of PCr_1v2, PCr_1v3, and PCr_1v4 were combined and averaged to produce the score 0.86357 ±0.18084. The Table 5.3 shows the stability score of each individual XAI method relative to all other XAI methods. Green and red highlights show the highest and lowest stability scores of 0.86357 ±0.18084 (Vanilla Saliency) & 0.78992 ±0.25374 (GradCAM++) respectively.

Figure 5.10 is a BW visualization of the distribution of individual pairwise comparisons (PCr scores) for each XAI method. Each BW plot show a distribution of 15570 pairwise comparisons (XAI outputs of 5190 test records • 3 pairwise comparisons).

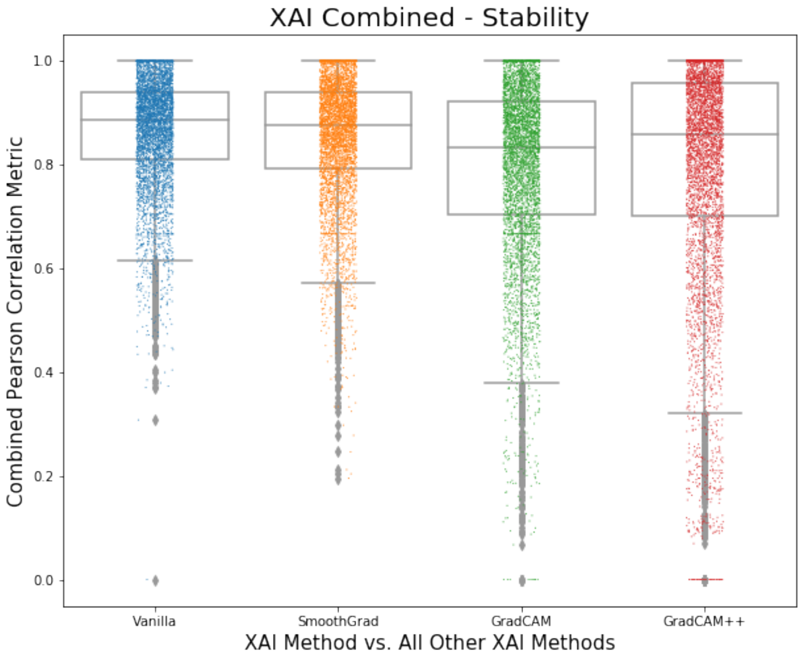


Figure 5.10: Visualization of Overall Stability Scores for Each XAI method - Box plot display the distribution statistics of the stability scores for each pairwise comparison (PCr score). The waterfall plot overlay, displays the relative density of PCr scores for each XAI method.

5.3.3 Novelty & Overlooked Learning

Comparing novel and overlooked features identified by the XAI methods provided insights into the performance of the underlying AI model and help determine its overall trustworthiness, as discussed in Section 5.2.3. TF-IDF scores were used to evaluate the utility of each individual feature.

Table 5.4 shows TF-IDF scores for novel features from a sample of records, for Vanilla Saliency XAI method only. Features are separated by classifications (SR, GSVT, AFIB, & SB) in columns and individual records in rows. Multiple features with different TF-IDF scores in the same record are grouped together. The feature with the lowest TF-IDF score for any given record is highlighted in red. In Table 5.4

class SB has no records with identified novel features, GSVT, AFIB and SR have 39, 541 and 3166 records with possible novel features identified respectively.

Table 5.4: Novel learned features according to Vanilla Saliency XAI - TF-IDF scores for each record with unique features identified by the XAI only, separated by true class labels for each record.

SR			GSVT			AFIB			SB		
TF-IDF	**Feature**	**Record**	**TF-IDF**	**Feature**	**Record**	**TF-IDF**	**Feature**	**Record**	**TF-IDF**	**Feature**	**Record**
0.00000	**0**	**0**	0.02404	0	0	0.02947	0	0			
0.41448	13	0	**0.00000**	**1**	**0**	**0.00000**	**1**	**0**			
0.44928	37	0	0.49167	4	0	0.29287	44	0			
0.39364	112	0	0.41003	66	0	0.25937	83	0			
0.41954	125	0	0.36227	240	0	0.27090	105	0			
0.41783	170	0	0.05129	311	0	0.25426	181	0			
0.42669	205	0	0.46485	358	0	0.23692	213	0			
0.40355	262	0	0.55741	393	0	0.22629	277	0			
0.46667	447	0	0.41071	206	1	0.23319	321	0			
0.43841	492	0	**0.37230**	**373**	**1**	0.28474	391	0			
	...			...			...				
0.05757	**0**	**3089**	**0.12371**	**172**	**32**	**0.72681**	**399**	**509**			
0.20731	**159**	**3109**	**0.07647**	**227**	**35**	**1.02178**	**167**	**516**			
0.19427	**217**	**3112**	0.22318	247	35	1.18730	406	516			
0.31597	**437**	**3155**	0.12830	38	36	**0.16823**	**495**	**520**			
0.20289	**142**	**3166**	**0.09018**	**349**	**36**	**0.36149**	**310**	**528**			
0.32054	433	3166	0.20732	489	39	**0.20867**	**215**	**541**			

Table 5.5 shows the TF-IDF scores for overlooked features that were never learned from a sample of records, for Vanilla Saliency XAI method only. The feature with the highest TF-IDF score for any given record is highlighted in green. In Table 5.5 GSVT, AFIB, SB, and SR have 37, 77, 95 and 1237 records with overlooked features identified respectively for the the Vanilla Saliency XAI.

Full tables for all four XAI methods are in Appendix A starting from Table A.4.

To help visualize the information shown in Tables 5.4 and 5.5 an example record (File102x010) was selected to show an example of novel and overlooked learning. An overlay of the record's ECG signal with the segmented processed XAI output was generated as shown in figure 5.11. This figure shows the ECG signal and locations where the AI paid the most attention in attempting to classify the ECG signal according to the Vanilla Saliency XAI method.

Table 5.5: Overlooked features according to Vanilla Saliency XAI - TF-IDF scores for each record with unique features identified by the SMEs only, separated by true class labels for each record.

SR			GSVT			AFIB			SB		
TF-IDF	**Feature**	**Record**	**TF-IDF**	**Feature**	**Record**	**TF-IDF**	**Feature**	**Record**	**TF-IDF**	**Feature**	**Record**
0.55669	0	0	0.00000	0	0	0.01668	0	0	0.00000	0	0
0.00000	1	0	0.78358	92	0	0.00000	1	0	0.42316	25	0
0.69965	**352**	**0**	**0.93825**	**300**	**0**	**0.52065**	**139**	**0**	**0.55256**	**84**	**0**
0.61241	**1**	**1**	0.10473	332	0	0.01281	303	0	0.38928	151	0
0.33074	259	1	0.41360	410	0	0.44762	327	0	0.47092	273	0
0.59535	**2**	**2**	0.36584	421	0	0.01362	106	1	0.36300	311	0
0.44770	260	2	0.09750	164	1	0.31893	125	1	0.34822	496	0
	...			...			...			...	
1.51324	**79**	**1191**	**0.37196**	**43**	**34**	0.00647	118	64	**0.21975**	**45**	**76**
0.42267	**213**	**1215**	0.30568	382	34	**1.10557**	**28**	**70**	**0.32337**	**365**	**78**
0.50582	**218**	**1220**	**0.24735**	**214**	**35**	**0.54806**	**32**	**74**	0.80170	37	95
0.54750	**53**	**1234**	0.06690	92	37	0.01010	284	77	**0.89217**	**73**	**95**
0.36557	**189**	**1237**	**0.22582**	**377**	**37**	**0.26672**	**304**	**77**	0.64702	233	95

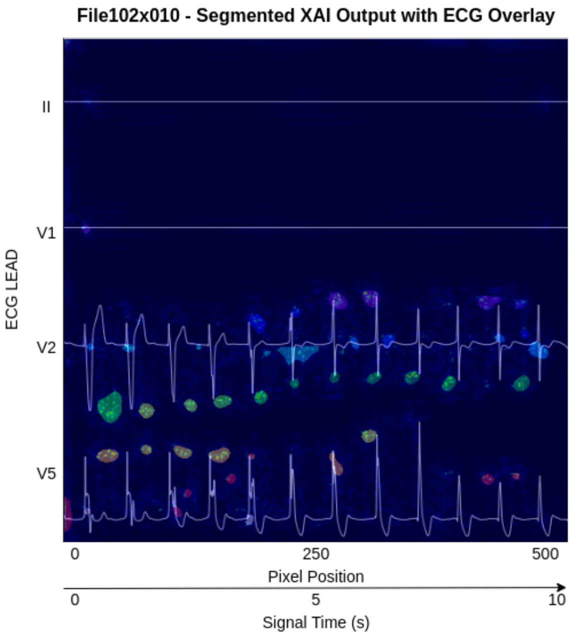


Figure 5.11: Visualization of XAI Method Output with ECG signal overlaid (coloured segments are used to represent distinct clusters only)

Table 5.6 is a a more detailed look at a single record (File102x010), from a list of records summarized in tables 5.4 & 5.5 and in Appendix A Tables A.4 & A.5. In addition to the details presented in other tables mentioned, the x-axis locations of the novel and overlooked features identified by the TF-IDF method are presented along with the filename for that record. The feature with the lowest TF-IDF score and overlooked feature with highest TF-IDF score were highlighted in red and green respectively, the colours are used to draw a visual connection to figures 5.12 & 5.13.

Table 5.6: Vanilla Saliency XAI Output - SR Class - List of Novel and Overlooked Features, the filename for the record and X-axis position on the record, of the feature of interest.

Novelty					Overlooked				
TF-IDF	**Feature**	**Record**	**Feat_x-axis**	**Filename**	**TF-IDF**	**Feature**	**Record**	**Feat_x-axis**	**Filename**
0.35053	124	255	21	file102x010	0.14599	10	10	107	file102x010
0.30658	**187**	**255**	**267**	**file102x010**	0.32728	134	10	219	file102x010
0.34672	415	255	472	file102x010	**0.33529**	**197**	**10**	**276**	**file102x010**
					0.32355	293	10	362	file102x010
					0.29873	470	10	71	file102x010

Figure 5.12 shows two parts, the figure on the left shows an ECG overlaid with Vanilla Saliency XAI segmented output and vertical lines marking the 3 novel features identified in table 5.6. The figure on the right only highlights the novel feature (red line) with the lowest TF-IDF score.

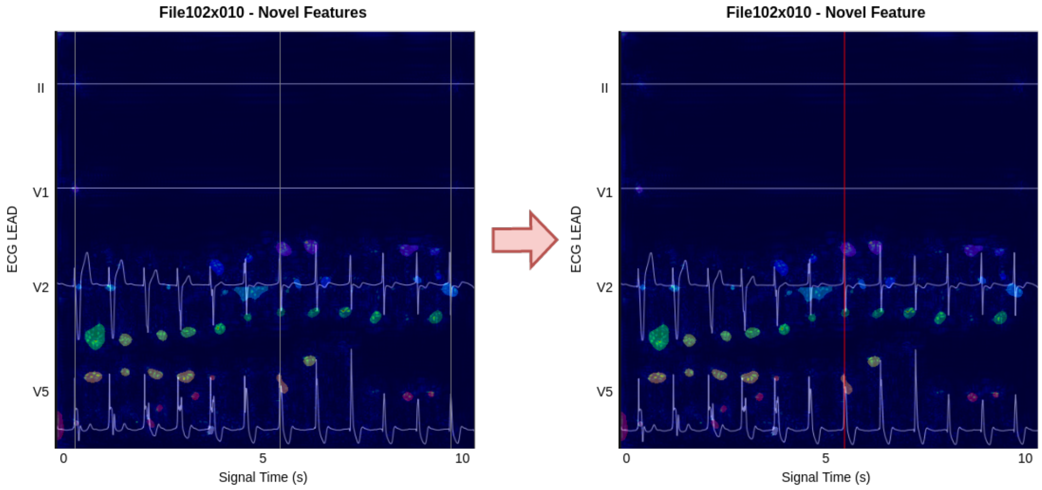


Figure 5.12: Visualization of Novel Features of Vanilla XAI Method Output Listed in Table 5.6. Image on the left presents all unique features, image on the right identifies feature with the lowest TF-IDF score (most useful novel feature).

Figure 5.13 also shows two parts, the figure on the left shows an ECG overlaid with Vanilla Saliency XAI segmented output and vertical lines marking the 5 overlooked features identified in table 5.6. The figure on the right only highlights the overlooked feature (green line) with the highest TF-IDF score.

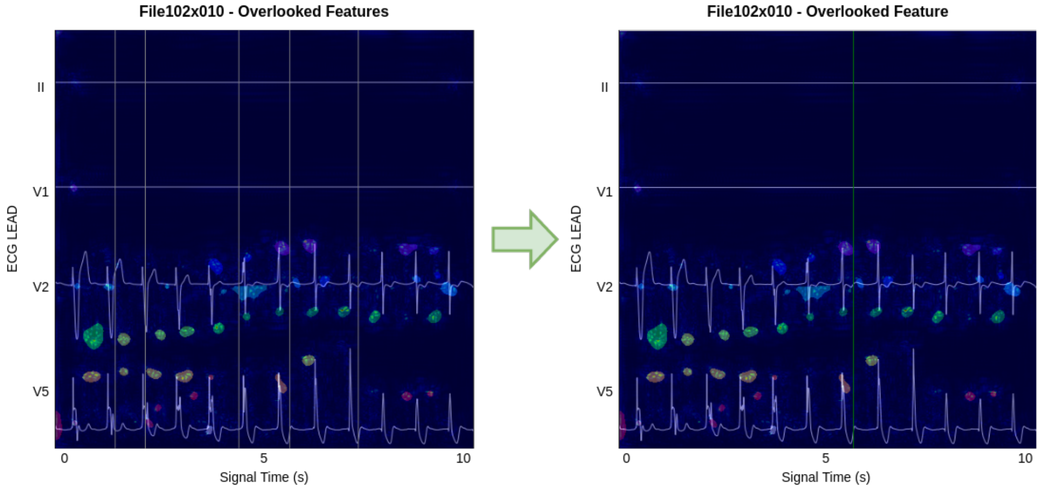


Figure 5.13: Visualization of Overlooked Features of Vanilla XAI Method Output Listed in Table 5.6 Image on the left presents all overlooked features, image on the right identifies feature with the highest TF-IDF score (most useful important overlooked feature).

Chapter 6

Quantifying Trust

6.1 Trust Score

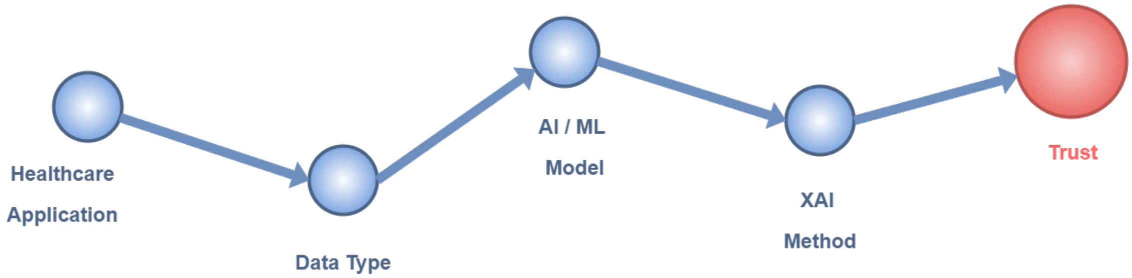


Figure 6.1: Visualization of overall Journey of the Thesis - The final step of identifying how trustworthy the XAI is, is determined in this section.

To improve user trustworthiness towards the AI, this thesis explores the impact of the explainability aspect of trust using a series of equations. This section implements quantification of trust using these equations.

6.2 Equations

The trust equations are reiterated in this section:

Overall trust equation:

$$T = \left(\left[\frac{\sum_i (w_x T_x)_i}{\sum_i w_i} \right] + w_v T_v + w_r T_r + w_f T_f \right) / \sum_n w_n \tag{2.4.1}$$

Explainability term from equation 2.4.1:

$$T_x = \frac{\sum(w_{sim} E_{sim},\ w_{stab} E_{stab})}{\sum w} \tag{2.4.6}$$

Similarity & Stability terms from equation 2.4.6:

$$E_{sim} = \frac{\sum(0.5 M_{Jaccard}, 0.5 M_{Hamming})}{1} ,\ 0 \leq M \leq 1 \tag{2.4.4}$$

$$E_{stab} = \frac{\sum(N_{PearsonCorr})}{1} ,\ 0 \leq N \leq 1 \tag{2.4.5}$$

Equations for Novel and Overlooked learning:

$$K_n = \frac{\Sigma(Record_n | XAI_{novelty})}{\Sigma(Record_n)} \tag{2.4.7}$$

$$R_u = \frac{\Sigma(Record_u | XAI_{overlooked})}{\Sigma(Record_u)} \tag{2.4.8}$$

Individual XAI's trustworthiness equation:

$$E_{XAI} = [T_x ,\ K_n ,\ R_u] ,\ 0 \leq T, K, R \leq 1 \tag{2.4.9}$$

6.3 Contribution to Trustworthiness

An example of the aforementioned equations as they apply to the Vanilla Saliency XAI method:

Vanilla Saliency:

$$\begin{aligned} E_{sim} &= \frac{\sum(\ 0.5 \cdot 0.64916, 0.5 \cdot 0.69470\ )}{1} \\ E_{sim} &= 0.67193 \end{aligned} \tag{6.3.1}$$

$$\begin{aligned} E_{stab} &= \frac{\sum(\ 1 \cdot 0.86357\ )}{1} \\ E_{stab} &= 0.86357 \end{aligned} \tag{6.3.2}$$

$$\begin{aligned} T_x &= \frac{\sum(\ 0.5 \cdot 0.67193,\ 0.5 \cdot 0.86357\ )}{1} \\ T_x &= 0.76775 \end{aligned} \tag{6.3.3}$$

$$\begin{aligned} K_n &= \frac{717|XAI_{overlooked}}{5190} \\ K_n &= 0.138 \end{aligned} \tag{6.3.4}$$

$$\begin{aligned} R_u &= \frac{348|XAI_{overlooked}}{5190} \\ R_u &= 0.067 \end{aligned} \tag{6.3.5}$$

$$E_{XAI} = [0.76775\ ,\ 0.138\ ,\ 0.067\ ] \tag{6.3.6}$$

A vector of values was created as shown in equations 6.3.6 to provide insight on the individual utility and trustworthiness of the Vanilla Saliency XAI method. The T_x term provided the trustworthiness measure for the XAI method, built upon similarity

(E_{sim}) and stability (E_{stab}) measures. K_n identified the novel Knowledge that the AI learned. R_u identified any recommendations the XAI had to improve the AI model by identifying important overlooked features.

Only the Vanilla Saliency XAI method was shown as a worked example. The full table of values for the trust equations from all XAI methods was calculated and presented in table 6.1.

Table 6.1: Trust Equation Values for XAI Methods - Presenting Similarity, Stability, Explainability (T_x), Novelty (K_n), & Overlooked (R_u) scores for all XAI methods, and E_xai XAI trustworthiness vector.

	Vanilla Saliency	**SmoothGrad**	**GradCAM**	**GradCAM++**
M_jaccard	0.64916	0.89502	0.58125	0.56321
M_hamming	0.69470	0.92105	0.62719	0.60477
E_similarity	0.67193	0.90803	0.60422	0.58399
N_pearson	0.86357	0.85111	0.79447	0.80262
E_stability	0.86357	0.85111	0.79447	0.80262
T_x	0.768	0.879	0.699	0.693
K_n	0.138	0.124	0.152	0.149
R_u	0.067	0.045	0.072	0.074
E_xai	**[0.768, 0.138, 0.067]**	**[0.879, 0.124, 0.045]**	**[0.699, 0.152, 0.072]**	**[0.693, 0.149, 0.074]**

Figure 6.2 shows the E_{xai} vector for each XAI method, with values taken from table 6.1 E_{xai} for each XAI method. The vectors present the XAI methods visualised relative to each other in terms of Explainability scores, as well as their novelty and overlooked learning scores. This presents insights into how XAI methods may related to each other in various dimensions.

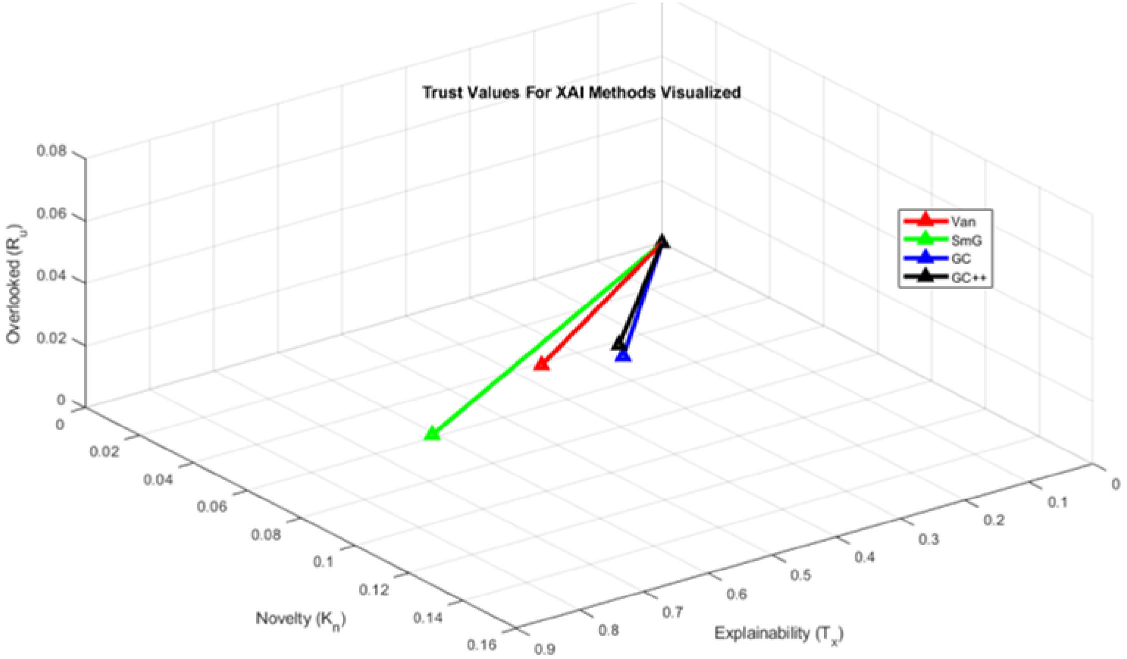


Figure 6.2: Visualization of E_xai vectors for each XAI method - Shows a visual representation of each XAI method on axis of Explainability, Novelty and Overlooked learning.

Table 6.1 also shows the E_{xai} vector for each XAI method. A vector of values that provides insights about explainability, novel and overlooked learning for a given XAI method. The results show that SmoothGrad had the highest explainability (0.87957) score among all the tested XAI methods, and identified the least novel and overlooked features in XAI's underlying learning. GradCAM and GradCAM++ had low similarity and stability scores compared to Vanilla or SmoothGrad and thus had low overall explainability scores (0.699 and 0.693 respectively). GradCAM and GradCAM++ provided more insights on novel learning and on overlooked learning and performed relatively similar to each other.

Chapter 7

Discussion

This section presents a discussion on the results provided in previous sections as well as identifying the limitations of the methods and the results and what those limitations mean.

7.1 Quantifying Trust Equations

Equation 2.4.9 in section 6 provided insights about the explanations generated by the XAI methods. Similarity and stability scores were combined together to produce T_x. A simple metric to help quickly identify how closely an XAI generated explanation resembles one produced by SMEs and other XAI methods. The novelty and overlooked learning scores K_n & R_u respectively, are metrics to rate the ability of the XAI method to identify what new features the AI learned or known features it missed in its learning. There is a pattern in the performance of the XAI methods shown by E_{xai} vectors in table 6.1. As the T_x value decreases, the likelihood of finding a novel or overlooked feature increases. This feature can be used to identify the XAI method best suited for specific stakeholders under specific conditions.

Applications that analyze the AI's learning, will find XAI methods with high K_n and R_u scores more useful to provide insights about the learning process of the AI. Compared to applications requiring an accurate and consistent explanation, the XAI method with higher T_x scores would be more preferable.

There is a need for validation of the trust scores E_{xai} for all the XAI methods by SMEs to determine how congruent the XAI generated explanation perceived qualities are relative to their scores. Thus a limitation of this study is that until the validation is performed, this thesis only presents a novel framework to objectively evaluate explanations and only proposes a relative performance ranking between the explanations generated by each XAI method.

7.2 XAI Methodologies

In section 5.2.1 clusters corresponding to areas with no ECG signal were discarded. This biases the outcome of similarity comparison metrics in favour of higher similarity between the XAI and SME. The reason for discarding the clusters was that while some of these clusters might represent incorrectly learned patterns by the underlying AI. It would require a significant effort to distinguish which clusters are a result of noise, actual learning, or artifacts introduced by the XAI methodology itself. Discarding this data for now results in a significant reduction in computational complexity. Since all the individual records are processed the same way, any impacts to the similarity comparison will apply to all records, minimizing skewness in the results.

In section 5.2.1 equation 2.3.5. There was an issue regarding the use of Jaccard similarity as a measure of comparison between datasets that some datasets could be arbitrarily larger or smaller relative to the paired dataset. There may be an arbitrarily large union between the two datasets, while the intersection may or may not be affected. Thus the results of the similarity comparison could be arbitrarily affected by the size of the compared datasets. This problem is overcome by ensuring that all pairs of dataset being compared are of identical sizes as shown in equation 5.2.1.

The individual features used as the cumulative feature vocabulary in section 5.2.3 are partially idealized. There is sometimes a shift at the starting point of the signal, as to what part of the ECG wave the signal starts from. This is due to the random nature of the sampling and stride windows and how they they randomly capture a sample from the overall record. This randomized shifting between records would result in desynchronized feature positions between samples. To minimize the impact of this shifting effect, the features were given a .5 pixel (50 samples) buffer around the feature position. This way the feature in the vocabulary for TF-IDF is idealized and comparable between records.

7.3 XAI Results

Similarity

In section 5.3.1 similarity metric results were plotted to visualize the differences between the performance of SmoothGrad XAI and GradCAM++ XAI. The expectation (see section 2.3.3) was that test records that are incorrectly classified by the underlying AI model will have lower similarity scores vs high similarity scores for correctly classified records. The assumption being that SME annotations are most likely dissimilar to what the AI actually paid attention to in incorrectly classified records.

The XAI methods that can produce high similarity scores for correct classifications, and low similarity scores for incorrect classifications provide a measure of accuracy of the underlying AI. This measure of accuracy is one of the concerns for all stakeholders (section 2.1.3) that affects the overall trustworthiness of the underlying AI.

The SmoothGrad XAI method produced higher similarity scores on the Hamming and Jaccard metrics for correctly classified records relative to incorrectly classified records. Figure 5.6 a & b both show SmoothGrad method's median similarity score is close to 1.0, and the red dots depicting incorrect classifications are mostly well below the median similarity score.

Additionally, the SmoothGrad XAI method had high similarity scores averaged over all classes (see figure 5.7). This means that SmoothGrad method was able to identify more features the underlying AI had learn that were closer to the thought pattern of the SME. Since the SME is considered the gold standard in explaining the classification decision, SmoothGrad XAI outputs are the closest to the gold standard. GradCAM++ produced the lowest average similarity scores across any classification, thus this method was furthest from identifying the underlying learning of the AI relative to the SME gold standard. Additionally, the GradCAM++ method was bad at distinguishing the similarity of correctly classified records to SME annotations vs. dissimilarity between SME annotations and the AI learning behind incorrectly classified records.

The Hamming and Jaccard similarity metrics are consistent in the rankings they produce, further corroborating the results. With two separate methods of measuring similarity between the XAI outputs and SME annotations, XAI method rankings remained the same.

The similarity scores effectively provide a measure of accuracy of the XAI methods. Some XAI methods depict the underlying learning of the AI model more closely what the human trainers might deem useful, some XAI methods may be depicting less informative components of the learned characteristics. The disparity in XAI methods' dissimilarity to the SME annotation, is more closely evaluated by looking at the similarity metrics of each individual record, as well as looking at the outputs for novelty and overlooked components in section 5.3.3, tables 5.4 & 5.5. But without looking at any other comparison methods like novelty, similarity begins to provide an objective form of evaluating the quality of an explanation generated by an XAI method.

Stability

Section 5.3.2 the XAI methods Vanilla Saliency and GradCAM were determined to be the most and least stable XAI methods respectively. On one hand the Vanilla Saliency score being the most stable makes somewhat sense, when considering the similarity scores. Since Vanilla Saliency was in the middle of the pack in the similarity rankings, it stands to reason that the score would be considered somewhat stable relative to the other XAI methods as well. What is counter intuitive is the fact that GradCAM which also gave a middling score in the similarity rankings somehow was the least stable method overall. Just from intuition, SmoothGrad (highest similarity score) or GradCAM++ (lowest similarity score) would have made a lot of sense as the least stable XAI method, but the data says otherwise.

This means that the stability of an XAI method's performance relative to all other XAI methods over any individual record is not a function of how closely the XAI method's output matches the SME annotation, rather it is a completely independent feature from the similarity score and thus noteworthy enough to be included in the overall trust measurements for a given XAI method.

In section 5.3.2 the XAI methods Vanilla Saliency and GradCAM were determined to be the most and least stable XAI methods respectively.Since Vanilla Saliency was in the middle of the pack in the similarity rankings, it would be reasonable to assume that the stability score should be higher relative to the other XAI methods as well. However, counter intuitively, GradCAM which also ranked middle of the pack in the similarity rankings, was the least stable method overall. Based on intuition alone, SmoothGrad (highest similarity score) and GradCAM++ (lowest similarity score) should have been the most and least stable XAI methods respectively, however the results of the stability metrics suggest otherwise.

Additionally, as seen in figure 5.9 the stability scores between Vanilla Saliency method generated outputs and SmoothGrad outputs was significantly higher than the rest of the pairwise comparisons for all classes. Subsequently, the overall stability scores for both Vanilla Saliency and SmoothGrad were relatively close together and higher than GradCAM and GradCAM++, which were also close to each other in value (see table 5.3). Given that SmoothGrad XAI method is a derivative of the Vanilla Saliency method and GradCAM++ is a derivative of GradCAM, these results are further support of the validity of the stability comparison method.

The stability of an XAI method's performance relative to all other XAI methods over any individual record is not a function of how closely the XAI method's output matches the SME annotation (similarity). Rather, it is a completely independent feature providing insight into the consistency of an XAI method's performance, and it affects the overall trustworthiness of the underlying AI.

Novelty

The TF-IDF methodology in section 5.2.3 applied to an individual record compares the importance of the features present in that record, using a full vocabulary of all features extracted from all the records. From section 5.3.3 when looking at the figure 5.12, the lowest scored TF-IDF feature (the least uniquely predictive feature) of that record (highlighted in red) is used to identify novel features.

In the common use of TF-IDF, a feature present in a record with a low TF-IDF score means it is not a unique enough feature to help identify that particular record; the feature is too common across all records. In the context of this thesis and identifying novel features, a low TF-IDF score means that this particular feature is present in many records which is interpreted to mean this feature is likely an actual novel feature identified by the underlying AI.

To demonstrate this in another terms, lets compare a series of arbitrary sentences:

“This is a unique sentence”,

“That is a common example”, and

“This is that example”.

The vocabulary from these sentences is as follows:

('a', 'common', 'example', 'is', 'sentence', 'that', 'This', 'unique')

Based on this vocabulary and how frequently each word/feature is used in all the sentences using the TF-IDF equation 2.3.9. Looking at sentence 1; “This is a unique sentence”; to a human reader it becomes obvious that the features “This” and “is”, are common among all the sentences. Which means they don't help uniquely identify the sentence, but they are an important feature to all the sentences because they structure the semantic meaning of the sentence.

Therefore, if the SME didn't identify any of these individual features as important when annotating, but an AI algorithm did identify them as meaningful features in its decision making criteria. Then it would be worth determining if the features that AI identified are informative and offer further insight.

For the purposes of the ECG signal evaluation in this thesis; the more common and not uniquely predictive a feature is the more likely it is to be present in ECG records of that class. The importance a feature plays to a record is measured by TF-IDF, and that is used to determine the utility of that feature in classification (SR, SB, AFIB, & GSVT) by the AI. Based on the rationale presented here, a feature with a low score is most likely a novel feature. However, to improve certainty in identifying novel features, the proposed method requires the SME to re-examine the signal and feature to determine the importance of the new insights that the AI might have gleaned. This method of determining novel features learned by the AI provides a better understanding of how the AI is improving upon the performance in the normal workflow. It provides specific insights into what the AI may have improved upon compared to SMEs.

Overlooked Learning

Similarly, from Section 5.3.3 when looking at the figure 5.13, the most predictive feature of that record (highlighted in green) is used to identify most important overlooked features. As discussed with the arbitrary sentences example above, in the sentence 1; "This is a unique sentence"; the features "unique" and "sentence" are present in only that one sentence but are very important in identifying that particular sentence. These features would have the largest TF-IDF score according to the equations 2.3.10. Since the AI did not identify these features, but an SME annotation exists, the XAI method can be used to identify the features that remain overlooked by the AI. Then a TF-IDF analysis determines which features are next most important features that were overlooked by the AI during training. Similar to the novel feature, these overlooked features need to be validated by the SME to determine their predictive importance.

If SME annotations exist for individual features of interest and AI has missed some of these features, as identified by the XAI method. A recursive training method can be implemented, that goes back to optimize the training of the AI, to ensure some number of the most informative features are learned by the AI. This methodology can help provide some degree of manual control over what features an AI learns to optimize performance where needed.

Chapter 8

Conclusion

This section provides future directions to explore and a summary of research contributions made by this thesis.

8.1 Summary

The goal of this research was to gain insight into how deep learning models make decisions so that user trust can be increased how the model identified features contribute to decisions made by the AI. This thesis provides a framework to evaluate the trustworthiness of AI model in medical applications with respect to the needs of the stakeholders involved. The thesis identified Explainability as one of the main constituent elements of trustworthiness and explored it in detail. Similarity, stability and novelty were identified as components to allow objective evaluation of explainability. This thesis hypothesized that measuring these components will create a framework to objectively evaluate explanations.

8.1.1 Components of Explainability

This thesis used a few existing distance metrics (Jaccard, Hamming, and Pearson Correlation) to measure similarity and stability between the explanations generated by XAI methods and the SMEs. These metrics inform the stakeholder of the level of similarity and stability an XAI method has relative to SMEs and other XAI methods.

Additionally, novel and overlooked learning are determined using TF-IDF scores. The scores measuring the presence of novel and overlooked learning are then combined with an average score measuring the similarity and stability of the XAI methods. These components together provide the user a repeatable and testable value to measure explainability, and provide insights about the explanations generated.

One of the insights about explanations generated is that the similarity and stability scores combine together to produce an inverse score to the novelty and overlooked learning scores. As discussed this information can be used to identify the best type of XAI methodology for the use case.

8.2 Future Directions

The XAI analysis provided measures of similarity and stability that allowed an exploration of the explainability aspect of trust in the underlying AI model. The analysis also provided some insight into the novel and overlooked learning by the underlying AI.

The proposed metrics and subsequent trust equations provide a basic framework to objectively evaluate XAI methods. However, more validation needs to be performed to increase confidence in the utility of the specific metrics and their ability to provide the objective evaluation of the XAI method.

In the immediate short term, one way to achieve this validation for similarity, and stability metrics is to use a synthetic ECG generator with predetermined features of diagnostic importance. Then utilize the various XAI methods on the outputs of VGG-16 based classifier model trained on the synthetic data. To determine the relative similarity and stability values of the XAI methods in a more rigorously controlled experiment.

Additionally, the utility of the novel and overlooked learning needs to be determined by a cardiology SME. The results of the outputs generated in this thesis can be provided to SMEs. They can determine if there is true predictive and/ or diagnostic value in the novel and overlooked learning outputs determined by the XAI analysis. Similarly, a number of SMEs should be presented with outputs from the trust equations 2.4.1 & 2.4.9 to determine if and how their trust in an AI model is impacted by the use of the equations.

Bibliography

[1] Canadian Medical Association, “Shaping the Future of Health and Medicine,” 2018.

[2] A. Tekkeşin, “Artificial Intelligence in Healthcare: Past, Present and Future,” *Anatolian journal of cardiology*, vol. 22, pp. 8–9, 2019.

[3] O. Asan, A. E. Bayrak, and A. Choudhury, “Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians,” *Journal of Medical Internet Research*, vol. 22, no. 6, pp. 1–7, 2020.

[4] M. Arnold, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, K. R. Varshney, R. K. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilovic, R. Nair, K. N. Ramamurthy, and A. Olteanu, “FactSheets: Increasing trust in AI services through supplier’s declarations of conformity,” *IBM Journal of Research and Development*, vol. 63, no. 4-5, pp. 1–13, 2019.

[5] S. Thiebes, S. Lins, and A. Sunyaev, “Trustworthy artificial intelligence,” *Electronic Markets*, vol. 31, no. 2, pp. 447–464, 2021.

[6] B. S. Zaunbrecher, S. Kowalewski, and M. Ziefle, “The willingness to adopt technologies: A cross-sectional study on the influence of technical self-efficacy

on acceptance," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8512 LNCS, no. PART 3, pp. 764–775, 2014.

[7] V. Tucci, J. Saary, and T. E. Doyle, "Factors influencing trust in medical artificial intelligence for healthcare professionals : a narrative review," no. November 2021, pp. 0–2, 2022.

[8] S. Castagno and M. Khalifa, "Perceptions of Artificial Intelligence Among Healthcare Staff: A Qualitative Survey Study," *Frontiers in Artificial Intelligence*, vol. 3, no. October, pp. 1–7, 2020.

[9] E. LaRosa and D. Danks, "Impacts on trust of healthcare ai," pp. 210–215, Association for Computing Machinery, Inc, 12 2018.

[10] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use," no. Ml, pp. 1–21, 2019.

[11] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," no. Ml, pp. 1–13, 2017.

[12] L. Rieger, C. Singh, W. J. Murdoch, and B. Yu, "Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge," *37th International Conference on Machine Learning, ICML 2020*, vol. PartF168147-11, pp. 8086–8096, 2020.

[13] J. Krause, A. Perer, and K. Ng, “Interacting with predictions: Visual inspection of black-box machine learning models,” *Conference on Human Factors in Computing Systems - Proceedings*, pp. 5686–5697, 2016.

[14] M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield, H. Khlaaf, J. Yang, H. Toner, R. Fong, T. Maharaj, P. W. Koh, S. Hooker, J. Leung, A. Trask, E. Bluemke, J. Lebensold, C. O'Keefe, M. Koren, T. Ryffel, J. Rubinovitz, T. Besiroglu, F. Carugati, J. Clark, P. Eckersley, S. de Haas, M. Johnson, B. Laurie, A. Ingerman, I. Krawczuk, A. Askell, R. Cammarota, A. Lohn, D. Krueger, C. Stix, P. Henderson, L. Graham, C. Prunkl, B. Martin, E. Seger, N. Zilberman, S. Ó. HÉigeartaigh, F. Kroeger, G. Sastry, R. Kagan, A. Weller, B. Tse, E. Barnes, A. Dafoe, P. Scharre, A. Herbert-Voss, M. Rasser, S. Sodhani, C. Flynn, T. K. Gilbert, L. Dyer, S. Khan, Y. Bengio, and M. Anderljung, “Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims,” 2020.

[15] D. Gunning and D. W. Aha, “DARPA's explainable artificial intelligence program,” *AI Magazine*, vol. 40, no. 2, pp. 44–58, 2019.

[16] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, “Explainable ai: A review of machine learning interpretability methods,” *Entropy*, vol. 23, no. 1, pp. 1–45, 2021.

[17] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila,

and F. Herrera, “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Information Fusion*, vol. 58, no. December 2019, pp. 82–115, 2020.

[18] Z. Zhang, J. Singh, U. Gadiraju, and A. Anand, “Dissonance between human and machine understanding,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, 2019.

[19] C. DeBrusk, E. Gürdeniz, S. Santhanam, and T. Schuermann, “Trusting the Mind of a Machine,” 2018.

[20] S. Mohseni, N. Zarei, and E. D. Ragan, “A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems,” *ACM Transactions on Interactive Intelligent Systems*, vol. 11, no. 3-4, pp. 1–45, 2021.

[21] A. Ferrario, M. Loi, and E. Viganò, “In AI We Trust Incrementally: a Multilayer Model of Trust to Analyze Human-Artificial Intelligence Interactions,” *Philosophy and Technology*, vol. 33, no. 3, pp. 523–539, 2020.

[22] H. Tao and J. Zhao, “Source Codes Oriented Software Trustworthiness Measure Based on Validation,” *Mathematical Problems in Engineering*, vol. 2018, 2018.

[23] S. Leijnen and F. van Veen, “The Neural Network Zoo,” *Proceedings*, vol. 47, no. 1, p. 9, 2020.

[24] R. Dipietro, N. Ahmidi, A. Malpani, M. Waldram, G. I. Lee, M. R. Lee, S. S. Vedula, and G. D. Hager, “Segmenting and classifying activities in robot-assisted surgery with recurrent neural networks,” *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 11, pp. 2005–2020, 2020.

[25] L. Dong, Y. Qin, L. Ya, C. Liang, H. Tinghui, H. Pinlin, Y. Jin, W. Youliang, C. Shu, and W. Tao, "Bayesian network analysis of open, laparoscopic, and robot-assisted radical cystectomy for bladder cancer," vol. 0, no. November, 2020.

[26] Z. Wang and A. M. Fey, "Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, no. 12, pp. 1959–1970, 2018.

[27] D. Ostler, M. Seibold, J. Fuchtmann, N. Samm, H. Feussner, D. Wilhelm, and N. Navab, "Acoustic signal analysis of instrument – tissue interaction for minimally invasive interventions," *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, no. 5, pp. 771–779, 2020.

[28] N. Sachdeva, M. Klopukh, R. St, C. William, and E. Hahn, "Using conditional generative adversarial networks to reduce the effects of latency in robotic telesurgery," *Journal of Robotic Surgery*, no. 0123456789, pp. 1–7, 2020.

[29] M. J. Fard, R. D. Ellis, and M. D. Klein, "Automated robot - assisted surgical skill evaluation : Predictive analytics approach," no. October 2016, pp. 1–10, 2018.

[30] S. Shorey, E. Ang, J. Yap, E. D. Ng, and S. T. Lau, "A Virtual Counseling Application Using Artificial Intelligence for Communication Skills Training in Nursing Education : Development Study Corresponding Author : Related Article :," vol. 21.

[31] M. Baig, N. Hua, E. Zhang, and R. Robinson, "Predicting Patients at Risk of 30-Day Unplanned Hospital Readmission," pp. 20–24, 2019.

[32] Y. Hwang, D. Yoon, E. K. Ahn, H. Hwang, and R. W. Park, "Provider risk factors for medication administration error alerts : analyses of a large-scale closed-loop medication administration system using RFID and barcode," no. July, pp. 1387–1396, 2016.

[33] T. Jilani, G. Housley, G. Figueredo, P.-s. Tang, J. Hatton, and D. Shaw, "International Journal of Medical Informatics Short and Long term predictions of Hospital emergency department attendances," *International Journal of Medical Informatics*, vol. 129, no. March 2018, pp. 167–174, 2019.

[34] C. Zhang, X. Xiao, and C. Wu, "Medical Fraud and Abuse Detection System Based on Machine Learning," 2020.

[35] M. Herland, R. A. Bauder, and T. M. Khoshgoftaar, "Approaches for identifying U . S . medicare fraud in provider claims data," no. May 2018, pp. 2–19, 2020.

[36] E. Lin, C.-h. Lin, and H.-y. Lane, "Precision Psychiatry Applications with Pharmacogenomics : Artificial Intelligence and Machine Learning Approaches,"

[37] A. K. Waljee, B. I. Wallace, S. Cohen-mekelburg, Y. Liu, B. Liu, K. Sauder, and R. W. Stidham, "Development and Validation of Machine Learning Models in Prediction of Remission in Patients With Moderate to Severe Crohn Disease," vol. 2, no. 5, pp. 1–10, 2019.

[38] R. Nimri, T. Battelino, L. M. Laffel, R. H. Slover, D. Schatz, S. A. Weinzimer, K. Dovc, T. Danne, and M. Phillip, "Insulin dose optimization using an automated artificial intelligence-based decision support system in youths with type 1 diabetes," vol. 26, no. September, 2020.

[39] D. Deshpande, J. G. Pasipanodya, S. G. Mpagama, P. Bendet, S. Srivastava, T. Koeuth, P. S. Lee, S. M. Bhavnani, P. G. Ambrose, G. Thwaites, S. K. Heysell, and T. Gumbo, "Levofloxacin Pharmacokinetics / Pharmacodynamics , Dosing , Susceptibility Breakpoints , and Artificial Intelligence in the Treatment of Multidrug-resistant Tuberculosis," vol. 67, no. Suppl 3, pp. 293–302, 2018.

[40] S. Harrer, P. Shah, B. Antony, and J. Hu, "Arti fi cial Intelligence for Clinical Trial Design," vol. 40, no. 8, pp. 577–591, 2019.

[41] M. Nagendran, Y. Chen, C. A. Lovejoy, A. C. Gordon, M. Komorowski, H. Harvey, E. J. Topol, J. P. A. Ioannidis, G. S. Collins, and M. Maruthappu, "Artificial intelligence versus clinicians : systematic review of design , reporting standards , and claims of deep learning studies," pp. 1–12, 2020.

[42] C. Xiao, E. Choi, and J. Sun, "Review Opportunities and challenges in developing deep learning models using electronic health records data : a systematic review," vol. 25, no. June, pp. 1419–1428, 2018.

[43] M. Gong, J. Feng, and Y. Xie, "Privacy-enhanced multi-party deep learning," *Neural Networks*, vol. 121, pp. 484–496, 2020.

[44] P. Data, "HHS Public Access," vol. 33, no. 5, pp. 887–893, 2020.

[45] Accenture, "Artificial Intelligence: Healthcare's New Nervous System," *Accenture Report*, pp. 1–8, 2017.

[46] A. Maslova, R. N. Ramirez, K. Ma, H. Schmutz, C. Wang, C. Fox, B. Ng, C. Benoist, and S. Mostafavi, "Deep learning of immune cell differentiation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 117, no. 41, pp. 25655–25666, 2020.

[47] T. J. Loftus, P. J. Tighe, A. C. Filiberto, P. A. Efron, S. C. Brakenridge, A. M. Mohr, P. Rashidi, G. R. Upchurch, and A. Bihorac, "Artificial Intelligence and Surgical Decision-making," *JAMA surgery*, vol. 155, no. 2, pp. 148–158, 2020.

[48] C. Krittanawong, K. W. Johnson, R. S. Rosenson, Z. Wang, M. Aydar, U. Baber, J. K. Min, W. H. W. Tang, J. L. Halperin, and S. M. Narayan, "Deep learning for cardiovascular medicine : a practical primer," 2019.

[49] J. Grischke, L. Johannsmeier, L. Eich, L. Griga, and S. Haddadin, "Dentronics: Towards robotics and artificial intelligence in dentistry," *Dental Materials*, vol. 36, no. 6, pp. 765–778, 2020.

[50] M. Phillips, H. Marsden, W. Jaffe, R. N. Matin, G. N. Wali, J. Greenhalgh, E. McGrath, R. James, E. Ladoyanni, A. Bewley, G. Argenziano, and I. Palamaras, "Assessment of Accuracy of an Artificial Intelligence Algorithm to Detect Melanoma in Images of Skin Lesions," *JAMA Network Open*, vol. 2, no. 10, pp. 1–12, 2019.

[51] E. Vorontsov and B. E. Sci, "Deep Learning : A Primer for," 2017.

[52] Y. Raita, T. Goto, M. K. Faridi, D. F. Brown, C. A. Camargo, and K. Hasegawa, "Emergency department triage prediction of clinical outcomes using machine learning models," *Critical Care*, vol. 23, no. 1, pp. 1–13, 2019.

[53] Y. Tian, J. Yang, M. Lan, and T. Zou, "Construction and analysis of a joint diagnosis model of random forest and artificial neural network for heart failure," *Aging*, vol. 12, no. 24, pp. 26221–26235, 2020.

[54] A. Ohsaka, "Artificial intelligence (ai) and hematological diseases: establishment of a peripheral blood convolutional neural network (cnn)-based digital morphology analysis system," *[Rinsho Ketsueki] The Japanese Journal of Clinical Hematology*, vol. 61, no. 5, pp. 564–569, 2020.

[55] C. Peruselli, L. De Panfilis, G. Gobber, M. Melo, and S. Tanzi, "Artificial intelligence and palliative care: opportunities and limitations.," *Recenti Progressi in Medicina*, vol. 111, no. 11, pp. 639–645, 2020.

[56] H. Liyanage, S. T. Liaw, J. Jonnagaddala, R. Schreiber, C. Kuziemsky, A. L. Terry, S. de Lusignan, and ..., "Artificial Intelligence in Primary Health Care: Perceptions, Issues, and Challenges: Primary Health Care Informatics Working Group Contribution to the . . . ," *Yearbook of medical . . .* , vol. 28, no. 1, pp. 41–46, 2019.

[57] O. Niel and P. Bastard, "Arti fi cial Intelligence in Nephrology : Core Concepts , Clinical Applications , and Perspectives," *American Journal of Kidney Diseases*, vol. 74, no. 6, pp. 803–810, 2019.

[58] N. M. Murray, M. Unberath, G. D. Hager, and F. K. Hui, "Artificial intelligence to diagnose ischemic stroke and identify large vessel occlusions: A systematic review," *Journal of NeuroInterventional Surgery*, vol. 12, no. 2, pp. 156–164, 2020.

[59] F. Nensa, A. Demircioglu, and C. Rischpler, "Artificial intelligence in nuclear medicine," *Journal of Nuclear Medicine*, vol. 60, no. 9, pp. 29S–37S, 2019.

[60] C. L. Curchoe and C. L. Bormann, "Artificial intelligence and machine learning for human reproduction and embryology presented at ASRM and ESHRE 2018," *Journal of Assisted Reproduction and Genetics*, vol. 36, no. 4, pp. 591–600, 2019.

[61] D. S. W. Ting, L. Peng, A. V. Varadarajan, P. A. Keane, P. M. Burlina, M. F. Chiang, L. Schmetterer, L. R. Pasquale, N. M. Bressler, D. R. Webster, M. Abramoff, and T. Y. Wong, "Progress in Retinal and Eye Research Deep learning in ophthalmology : The technical and clinical considerations," *Progress in Retinal and Eye Research*, vol. 72, no. April, p. 100759, 2019.

[62] S. Sunny, A. B. Id, B. L. James, D. Balaji, N. V. Aparna, M. H. Rana, P. Gurpur, A. Skandarajah, M. D. Ambrosio, R. D. Ramanjinappa, S. P. Mohan, N. Raghavan, U. Kandasarma, N. Sangeetha, S. Raghavan, N. Hedne, F. Koch, D. A. Fletcher, S. Selvam, M. Kollegal, P. B. N, L. Ladic, A. Suresh, H. J. Pandya, and A. K. Id, "RESEARCH ARTICLE A smart tele-cytology point-of-care platform for oral cancer screening," pp. 1–16, 2019.

[63] H. Liang, B. Y. Tsui, H. Ni, C. Valentim, S. L. Baxter, G. Liu, W. Cai, D. S. Kermany, X. Sun, J. Chen, *et al.*, "Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence," *Nature medicine*, vol. 25, no. 3, pp. 433–438, 2019.

[64] S. H. Chae, Y. Kim, K.-S. Lee, and H.-S. Park, "Development and clinical evaluation of a web-based upper limb home rehabilitation system using a smartwatch and machine learning model for chronic stroke survivors: prospective comparative study," *JMIR mHealth and uHealth*, vol. 8, no. 7, p. e17216, 2020.

[65] Z. S. Wong, J. Zhou, and Q. Zhang, "Artificial intelligence for infectious disease big data analytics," *Infection, disease & health*, vol. 24, no. 1, pp. 44–48, 2019.

[66] V. Y. Londhe and B. Bhasin, "Artificial intelligence and its potential in oncology," *Drug Discovery Today*, vol. 24, no. 1, pp. 228–232, 2019.

[67] U. Ferizi, S. Honig, and G. Chang, "Artificial intelligence, osteoporosis and fragility fractures," *Current opinion in rheumatology*, vol. 31, no. 4, p. 368, 2019.

[68] H. Li and Y. Guan, "Deepsleep convolutional neural network allows accurate and fast detection of sleep arousal," *Communications biology*, vol. 4, no. 1, pp. 1–11, 2021.

[69] H. A. Elmarakeby, J. Hwang, R. Arafeh, J. Crowdis, S. Gang, D. Liu, S. H. AlDubayan, K. Salari, S. Kregel, C. Richter, *et al.*, "Biologically informed deep neural network for prostate cancer discovery," *Nature*, vol. 598, no. 7880, pp. 348–352, 2021.

[70] L. Li, M. Doroslovacki, and M. H. Loew, “Approximating the gradient of cross-entropy loss function,” *IEEE Access*, vol. 8, pp. 111626–111635, 2020.

[71] F. Chollet *et al.*, “Keras,” 2015.

[72] S. Ruder, “An overview of gradient descent optimization algorithms,” 9 2016.

[73] R. Caruana, P. Koch, Y. Lou, J. Gehrke, and M. Sturm, “Intelligible Models for HealthCare : Predicting Pneumonia Risk and Hospital 30-day Readmission,” pp. 1721–1730.

[74] B. Kim, E. Glassman, B. Johnson, B. Kim, E. Glassman, B. Johnson, and J. Shah, “Computer Science and Artificial Intelligence Laboratory Technical Report iBCM : Interactive Bayesian Case Model Empowering Humans via Intuitive Interaction,” 2015.

[75] S. Tan, G. Hooker, and M. T. Wells, “Tree Space Prototypes : Another Look at Making Tree Ensembles Interpretable,”

[76] H. Deng, “Interpreting Tree Ensembles with inTrees,”

[77] S. H. Welling, H. H. F. Refsgaard, P. B. Brockhoff, and H. Line, “Forest Floor Visualizations of Random Forests,” 2016.

[78] N. F. Rajani and R. J. Mooney, “Stacking With Auxiliary Features for Visual Question Answering,” pp. 2217–2226, 2018.

[79] N. Barakat and J. Diederich, “Eclectic Rule-Extraction from Support Vector Machines,” vol. 2, no. 1, pp. 59–62, 2005.

[80] P. Sollich, “Bayesian Methods for Support Vector Machines :,” pp. 21–52, 2002.

[81] W. Landecker, M. D. Thomure, M. A. Bettencourt, M. Mitchell, G. T. Kenyon, and S. P. Brumby, "Interpreting Individual Classifications of Hierarchical Networks," pp. 32–38, 2013.

[82] L. Rosenbaum, G. Hinselmann, A. Jahn, and A. Zell, "Interpreting linear support vector machine models with heat map molecule coloring," pp. 1–12, 2011.

[83] M. G. A. T. Kathirvalavakumar, "Reverse Engineering the Neural Networks for Rule Extraction in Classification Problems," pp. 131–150, 2012.

[84] M. Wu, M. C. Hughes, S. Parbhoo, M. Zazzi, V. Roth, and F. Doshi-velez, "Beyond Sparsity : Tree Regularization of Deep Models for Interpretability," pp. 1670–1678, 2018.

[85] J. Dean, "Distilling the Knowledge in a Neural Network," pp. 1–9.

[86] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: removing noise by adding noise," 2017.

[87] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *2nd International Conference on Learning Representations, ICLR 2014 - Workshop Track Proceedings*, pp. 1–8, 2014.

[88] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization," *Revista do Hospital das Cl??nicas*, vol. 17, pp. 331–336, 2016.

[89] "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," *Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018*, vol. 2018-January, pp. 839–847, 2018.

[90] M. M. Deza and E. Deza, *Encyclopedia of Distances.* 2016.

[91] T. Kato, I. Shimizu, and T. Pajdla, "Ipsj transactions on computer vision and applications selecting image pairs for sfm by introducing jaccard similarity," 2017.

[92] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, pp. 1–13, 2019.

[93] I. Mollas, N. Bassiliades, and G. Tsoumakas, "Altruist: Argumentative Explanations through Local Interpretations of Predictive Models," 2020.

[94] G. I. Heald, "A mathematical description of Trust," No. December 2019, pp. 11–16, 2020.

[95] A. Salehi-abari and T. White, "A Mathematical Analysis of Computational Trust Models with The Introduction of Con-man Agents ," *Computer*, pp. 1–28, 2010.

[96] H. Jiang, B. Kim, M. Gupta, and M. Y. Guan, "To trust or not to trust a classifier," *Advances in Neural Information Processing Systems*, vol. 2018-Decem, no. Nips, pp. 5541–5552, 2018.

[97] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet

Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[98] J. Zheng, "Chapmanecg," Jun 2019.

[99] J. Zheng, J. Zhang, S. Danioko, H. Yao, H. Guo, and C. Rakovski, "A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients," *Scientific Data*, vol. 7, 12 2020.

[100] "The impact of the mit-bih arrhythmia database history, lessons learned, and its influence on current and future databases."

[101] A. L. Goldberger, L. A. N. Amaral, . L. Glass, J. M. Hausdorff, . Plamen, C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and . H. E. Stanley, "Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals," 2000.

[102] M. Naz, J. H. Shah, M. A. Khan, M. Sharif, M. Raza, and R. Damaševičius, "From ecg signals to images: a transformation based approach for deep learning," *PeerJ Computer Science*, vol. 7, pp. 1–18, 2021.

[103] T. Pereira, C. Ding, K. Gadhoumi, N. Tran, R. A. Colorado, K. Meisel, and X. Hu, "Deep learning approaches for plethysmography signal quality assessment in the presence of atrial fibrillation," *Physiological Measurement*, vol. 40, 12 2019.

[104] M. Sallem, A. Ghrissi, A. Saadaoui, and V. Zarzoso, "Detection of cardiac arrhythmias from varied length multichannel electrocardiogram recordings using

deep convolutional neural networks detection of cardiac ar-rhythmias from varied length multichannel electrocardiogram recordings using deep convolu-tional neural networks. computing in cardology,"

[105] S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000. benchmark analysis of representative deep neural network architectures,"

[106] Y. Wen, L. Chen, Y. Deng, and C. Zhou, "Rethinking pre-training on medical imaging," *Journal of Visual Communication and Image Representation*, vol. 78, 7 2021.

[107] X. Zhang, J. Zou, K. He, and J. Sun, "Accelerating very deep convolutional networks for classification and detection," 5 2015.

[108] E. Gordon-Rodriguez, G. Loaiza-Ganem, G. Pleiss, and J. P. Cunningham, "Uses and abuses of the cross-entropy loss: Case studies in modern deep learning."

[109] K. Janocha and W. M. Czarnecki, "On loss functions for deep neural networks in classification," 2 2017.

[110] L. N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay," 3 2018.

[111] A. Ng, "Andrew ng notes collection," 8 2019.

[112] J. Quinonero-Candela, *Dataset shift in machine learning.* MIT Press, 2009.

[113] S. Saeb, L. Lonini, A. Jayaraman, D. C. Mohr, and K. P. Kording, "The need to approximate the use-case in clinical machine learning," *GigaScience*, vol. 6, pp. 1–9, 5 2017.

[114] R. Kotikalapudi and contributors, "keras-vis." `https://github.com/raghakot/keras-vis`, 2017.

[115] E. Arias-Castro and D. L. Donoho, "Does median filtering truly preserve edges better than linear filtering?," *Annals of Statistics*, vol. 37, pp. 1172–1206, 6 2009.

[116] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[117] S. L. Bangare, A. Dubal, P. S. Bangare, and S. T. Patil, "Reviewing otsu's method for image thresholding," *International Journal of Applied Engineering Research*, vol. 10, pp. 21777–21783, 2015.

Appendix A

Your Appendix

1 include a list of all snomed codes - look up the merged_diagnosis.csv

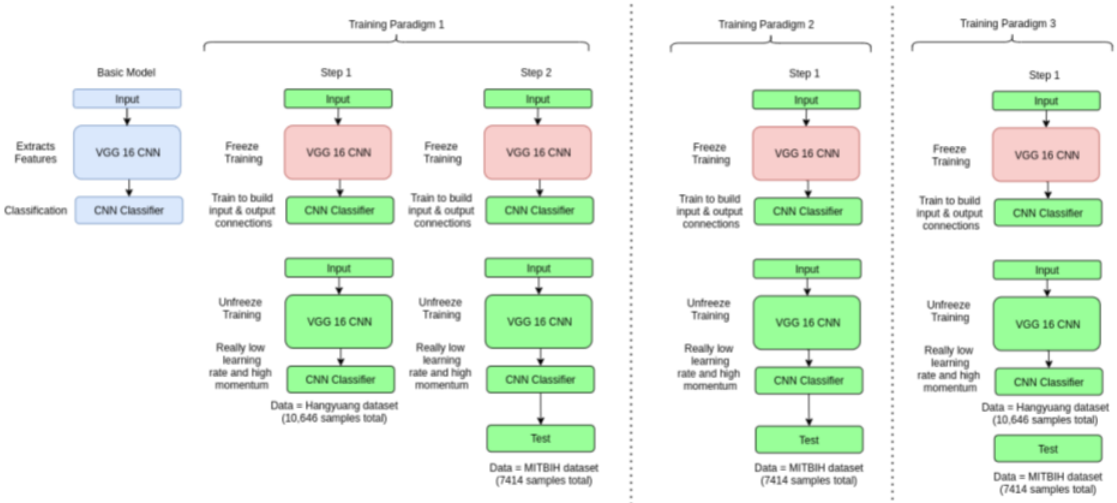


Figure A.1: Visualization of all paradigms for training the AI model used to identify the ideal training paradigm and accompanying parameters.

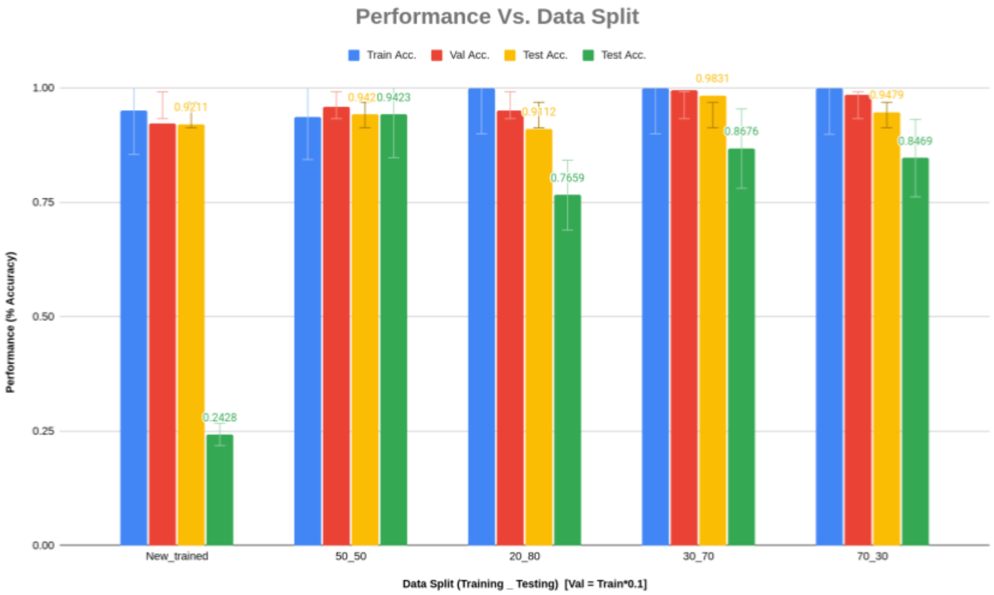


Figure A.2: Visualization of impact of data split between training, validation and testing used to identify the ideal data split.

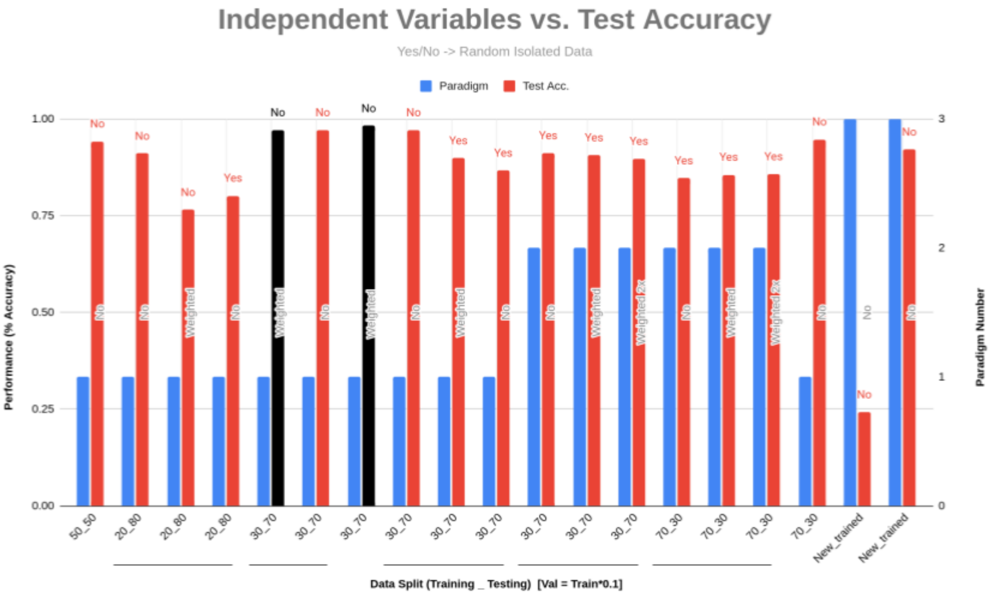


Figure A.3: Visualization of Independent Variables vs. Accuracy for models tested with all paradigms for training the AI model used to identify the ideal training paradigm and accompanying parameters.

Table A.1: Raw Data to accompany Figure A.6, Independent Variables vs. Accuracy for models tested with all paradigms for training the AI model used to identify the ideal training paradigm and accompanying parameters.

Paradigm	**Model Label**	**Pre-trained Model**	**Subj. Iso.**	**Rand. Iso.**	**Balancing**	**Data Split**	**Train Acc.**	**Val Acc.**	**Test Acc.**
1	50_50	pretrained	No	No	No	45/5/50	0.9375	0.9596	0.9423
1	20_80	pretrained	No	No	No	18/2/80	0.8791	0.8523	0.9112
1	20_80	pretrained	Yes	No	Weighted	22.5/2.5/75	0.981	0.95	0.7659
1	20_80	pretrained	Yes	Yes	No	22.5/2.5/75	1	0.8635	0.8021
1	30_70	None	No	No	Weighted	27/3/70	1	0.955	0.9715
1	30_70	None	No	No	No	27/3/70	0.9669	0.965	0.9714
1	30_70	pretrained	No	No	Weighted	27/3/70	0.985	0.9787	0.9831
1	30_70	pretrained	No	No	No	27/3/70	0.985	0.9787	0.9718
1	30_70	pretrained	Yes	Yes	Weighted	27/3/70	0.901	0.9281	0.8994
1	30_70	pretrained	Yes	Yes	No	27/3/70	0.9569	0.9679	0.8676
2	30_70	None	Yes	Yes	No	27/3/70	0.9837	0.9801	0.9121
2	30_70	None	Yes	Yes	Weighted	27/3/70	0.9875	0.9956	0.9066
2	30_70	None	Yes	Yes	Weighted 2x	27/3/70	0.9931	0.9956	0.8974
2	70_30	None	Yes	Yes	No	63/7/30	0.9937	0.9859	0.8469
2	70_30	None	Yes	Yes	Weighted	63/7/30	0.9987	0.9806	0.8565
2	70_30	None	Yes	Yes	Weighted 2x	63/7/30	0.985	0.9772	0.8585
1	70_30	pretrained	No	No	No	63/7/30	0.9406	0.95	0.9479
3	New_trained	pretrained	No	No	No	0/0/100	-	-	0.2428
3	New_trained	pretrained	No	No	No	70/10/20	0.95	0.9218	0.9211

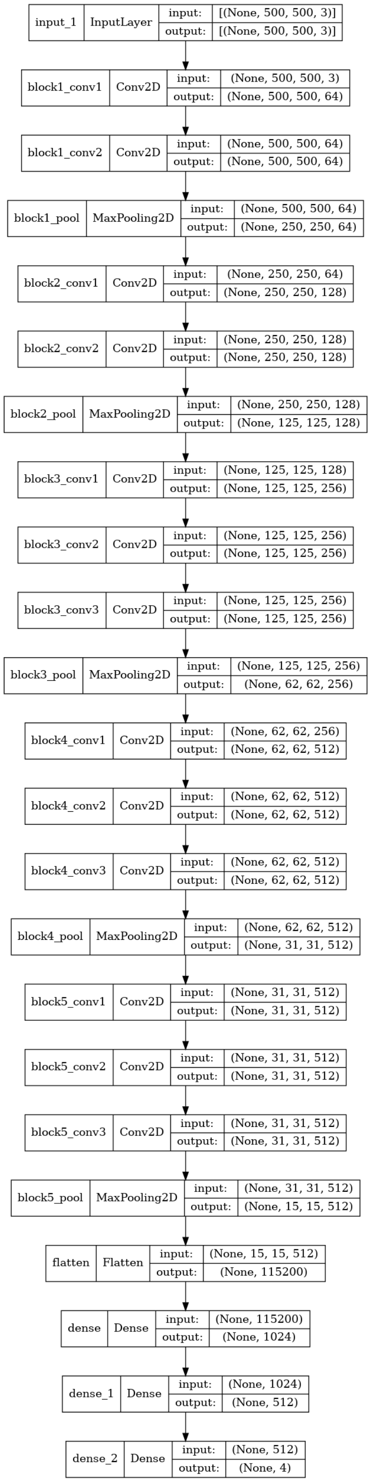

input_1
InputLayer
input: [(None, 500, 500, 3)]
output: [(None, 500, 500, 3)]
block1_conv1
Conv2D
input: (None, 500, 500, 3)
output: (None, 500, 500, 64)
block1_conv2
Conv2D
input: (None, 500, 500, 64)
output: (None, 500, 500, 64)
block1_pool
MaxPooling2D
input: (None, 500, 500, 64)
output: (None, 250, 250, 64)
block2_conv1
Conv2D
input: (None, 250, 250, 64)
output: (None, 250, 250, 128)
block2_conv2
Conv2D
input: (None, 250, 250, 128)
output: (None, 250, 250, 128)
block2_pool
MaxPooling2D
input: (None, 250, 250, 128)
output: (None, 125, 125, 128)
block3_conv1
Conv2D
input: (None, 125, 125, 128)
output: (None, 125, 125, 256)
block3_conv2
Conv2D
input: (None, 125, 125, 256)
output: (None, 125, 125, 256)
block3_conv3
Conv2D
input: (None, 125, 125, 256)
output: (None, 125, 125, 256)
block3_pool
MaxPooling2D
input: (None, 125, 125, 256)
output: (None, 62, 62, 256)
block4_conv1
Conv2D
input: (None, 62, 62, 256)
output: (None, 62, 62, 512)
block4_conv2
Conv2D
input: (None, 62, 62, 512)
output: (None, 62, 62, 512)
block4_conv3
Conv2D
input: (None, 62, 62, 512)
output: (None, 62, 62, 512)
block4_pool
MaxPooling2D
input: (None, 62, 62, 512)
output: (None, 31, 31, 512)
block5_conv1
Conv2D
input: (None, 31, 31, 512)
output: (None, 31, 31, 512)
block5_conv2
Conv2D
input: (None, 31, 31, 512)
output: (None, 31, 31, 512)
block5_conv3
Conv2D
input: (None, 31, 31, 512)
output: (None, 31, 31, 512)
block5_pool
MaxPooling2D
input: (None, 31, 31, 512)
output: (None, 15, 15, 512)
flatten
Flatten
input: (None, 15, 15, 512)
output: (None, 115200)
dense
Dense
input: (None, 115200)
output: (None, 1024)
dense_1
Dense
input: (None, 1024)
output: (None, 512)
dense_2
Dense
input: (None, 512)
output: (None, 4)

Figure A.4: Summary of the VGG-16 based deep learning model used for classification.

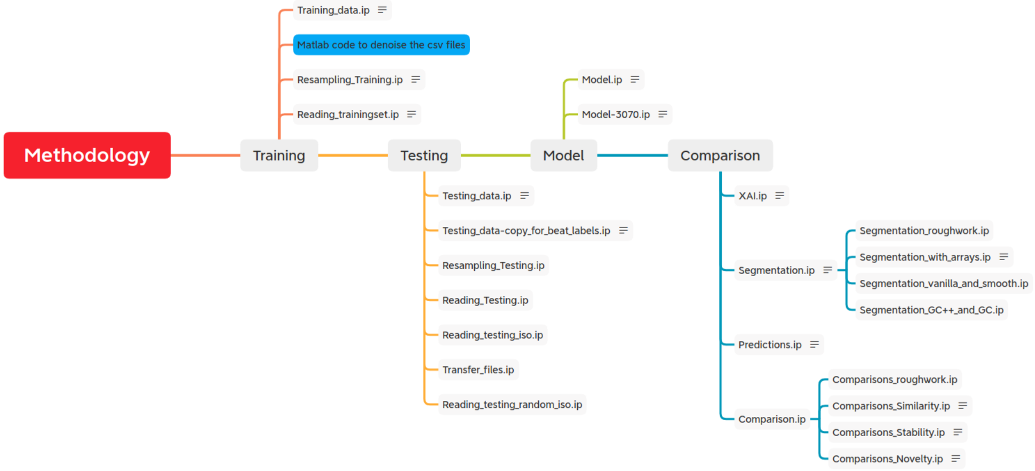


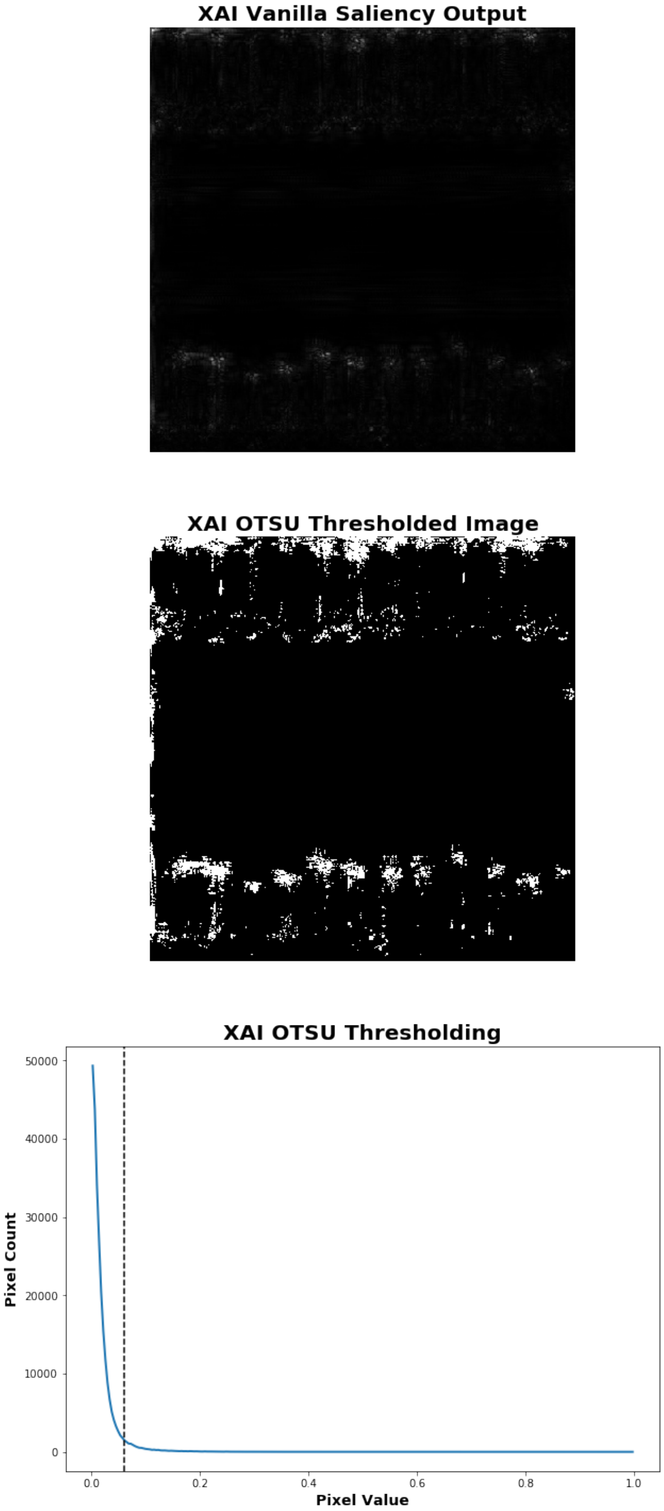
Figure A.5: Flowchart of the full methodology and order of implementation of accompanying files provided in the associated github for replicating work presented in this thesis.

AI models with related and compatible XAI methods

AI Models	**Tran.**	**AI Data Type**	**XAI Methods**	**Vis. Type**	**LO/GL**	**MA/MS**	**Source**
Logistic Regression	IN		-	-	-	-	ref
Linear Regression	IN		-	-	-	-	ref
Decision Trees	IN		-	-	-	-	ref
K-NN	IN		-	-	-	-	ref
Rule Based Learners	IN		-	-	-	-	ref
General Additive Model	IN	TAB	GA2M	TAB	GL	MS	ref
Bayesian Model	IN	GRAPH, TAB, TEXT	iBCM	GRAPH, TAB,TEXT	GL	MS	ref
Ensemble Methods & Decision Trees	PH	TAB	defragTrees	TAB	GL	MS	ref
Ensemble Methods & Decision Trees	PH	TAB	inTrees	TAB	GL	MS	ref
Random Forest	PH	TAB	Forest Floor	TAB, IMG	Both	MS	ref
CNN, RNN, LSTM, GRU	PH	IMG, TAB	SWAF (Stacking With Auxiliary Features)	TEXT	LO	MS	ref
SVM	PH	TAB	hybrid Rule-Extraction	TAB	GL	MS	ref
SVM	PH	TAB	Bayesian Method	TAB	GL	MS	ref
SVM	PH	IMG	Contribution Propagation	IMG	LO	MS	ref
Linear SVM	PH	IMG, TAB	Heatmap coloring viewer	TAB, IMG	GL	MS	ref
NN, ANN, FF	PH	TAB	RxREN (Rule Extraction by Reverese Engineering)	TAB	GL	MS	ref
NN, Perceptron, RNN w/ GRU,	PH	TAB	Tree Regularization	TAB	GL	MS	ref
NN, DCN,	PH	IMG	Distilling ensemble	IMG	GL	MS	ref

NN	PH			TAB	LO	MS	ref
DCN, CNN, MLP	PH	IMG	Assess Parameter Sensitivity	IMG	LO	MS	ref
NN, AE, FF, RNN, LSTM, GRU, CNN	PH	TEXT	rcNN	TEXT	GL	MS	ref
DCN	PH	IMG	DkNN (Deepest K-NN)	IMG, TEXT	Both	MS	ref
CNN, DCN	PH	IMG, TAB, TEXT	CDEP (Contextual Decomposition)	TAB, IMG, TEXT	GL	MS	ref
CNN, DCN	PH	IMG, TEXT	Sensitivity Analysis	IMG, TEXT	Both	MS	ref
CNN	PH	IMG	LRP	IMG	Both	MA	ref
RNN, LSTM	PH	TEXT	Auto Rule Extraction	TEXT	GL	MS	ref
BiRNN, LSTM, GRU	PH	TEXT	LRP for LSTM	TEXT	LO	MS	ref
RNN, LSTM, GRU	PH	TEXT	VisRNN	IMG,TEXT	GL	MS	ref
RNN, LSTM, GRU	PH	TAB	RETAIN (REverse Time AttentIoN model)	TAB	GL	MS	ref
Reg., SVM, Decision Tree, RF	PH	TAB	QII (Quantitative Input Influence)	TAB	LO	MA	ref
NN, SVM, Ensemble Methods	PH	TAB	DSA (Data-based SA), MSA (Monte-Carlo SA), CSA (Cluster-based SA)	TAB	LO	MA	ref
DCN, CNN, NN	PH	IMG	DeepSHAP	IMG	LO	MA	ref
SVM, NN, RF	PH	TAB	ASTRID	TAB	LO	MA	ref
NN, Decision Trees, Clustering, Ensemble Methods, Stat. Models	PH	TAB	ALEplot (Accumulated Local Effects)	TAB	LO	MA	ref
ANY, Classification	PH	IMG, TAB, TEXT	LIME	TAB, IMG, TEXT	Both	MA	ref
ANY, Classification	PH	TAB	LORE (Local Rule-based Explanation)	TAB	LO	MA	ref
ANY, Decision Trees, Random Forest	PH	TAB	Prospector	TAB	LO	MA	ref
CNN, DCN, NN,	PH	IMG	Realtime Saliency	IMG	GL	MA	ref
ANY, Classification	PH	TAB, IMG	DarkSight	TAB, IMG	GL	MA	ref
MC, Classification Trees	IN	TAB	Bayesian averaging over decision trees	TAB	GL	MS	ref
SVM, Logistic Regression, LDA	IN	TAB	SPDA (Sparsed Penalized Discriminant Analysis)	TEXT	GL	MS	ref
DBN, AE, DAE, BM, RBM	PH	IMG	Activation Maximization	IMG	LO	MA	ref
DCN, CNN, DBN, DN,	PH	IMG	Gradient-based Saliency Maps	IMG	LO	MA	ref
Classificaion, LDA, SVM	IN	ANY	BCM (Bayesian Case Model)	ANY	GL	MS	ref
DN	PH	IMG	DeConvolutional Nets	IMG	LO	MA	ref
CNN	PH	IMG	Guided Backprop	IMG	LO	MA	ref
BN, MCMC	IN	TAB	Bayes Rule Lists	TAB	GL	MS	ref
CNN	PH	IMG	CAM	IMG	LO	MA	ref
DCN, Ensemble Methods	PH	TAB, IMG, TEXT	SHAP (SHapley Additive exPlanations)	TAB, IMG, TEXT	Both	MA	ref
DCN, CNN	PH	IMG	Grad-CAM	IMG	LO	MA	ref

DCN	PH	IMG	PDA (Prediction Difference Analysis)	IMG	LO	MA	ref
DCN, MLP	PH	IMG	Deep Taylor Decomposition	IMG	LO	MA	ref
DCN	PH	IMG	Sensitivity-n (Gradient-based attribution method)	IMG	LO	MA	ref
Linear CNN, DN	PH	IMG	IG (Integrated Gradients)	IMG	LO	MA	ref
MLP, CNN, DCN	PH	IMG	PatternNet and PatternAttribution	IMG	LO	MA	ref
Classification, Regression	PH	IMG	TCAV (Testing with Concept Activation Vectors)	IMG	GL	MA	ref
CNN, DCN, Image Classification NN	PH	IMG	RISE	IMG	LO	MA	ref
DCN, CNN	PH	IMG	Grad-CAM++	IMG	LO	MA	ref
DCN, CNN	PH	IMG	(IRT & OSFT) Interpretability Randomization Test and One-Shot Feature Testing	IMG	LO	MA	ref
DCN, CNN	PH	IMG	SR (Salient Relevance) map	IMG	LO	MA	ref
DCN, CNN	PH	IMG	Spectral Relevance Analysis	IMG	GL	MA	ref
CNN	PH	IMG	GAM (Global Attribution Mapping)	IMG	GL	MA	ref
CNN	PH	TAB, IMG	ACE (Automatic Concept-based Explanations)	IMG	GL	MA	ref
VAE, AE, DCN	PH	IMG	CaCE (Causal Concept Effect)	IMG	GL	MA	ref
DCN	IN	IMG	NAMs (Neural Additive Models)	IMG	GL	MS	ref
CNN, DCN	PH	IMG	SG (SmoothGrad)	IMG	LO	MA	ref


XAI Vanilla Saliency Output

XAI OTSU Thresholded Image

XAI OTSU Thresholding

Figure A.6: Visualization of Otsu binary thresholding (dashed line at 0.060543224), all pixels are .

OTSU threshold is determined by creating a histogram representation of pixel values for an image and then separating the histogram into two classes using a threshold to minimize intra-class intensity variance, or equivalently, by maximizing inter-class variance [117]. This method is an analogue of linear discriminant analysis (LDA) or equivalent to K-means clustering on the intensity histogram to determine the optimal threshold to create a binary image.

Table A.3: Similarity Results - Full Dataset

XAI	**Metric**	**Cor.**	**AFIB**	**GSVT**	**SB**	**SR**	**Combined**
Vanilla	**Hamming**	**Y**	count 562.000000	count 79.000000	count 119.000000	count 4119.000000	count 4879.000000
			mean 0.678622	mean 0.550406	mean 0.614651	mean 0.706830	mean 0.698799
			std 0.214634	std 0.229354	std 0.243665	std 0.199468	std 0.204438
			min 0.000000	min 0.076900	min 0.000000	min 0.000000	min 0.000000
			25% 0.538500	25% 0.348450	25% 0.454500	25% 0.583300	25% 0.571400
			50% 0.699100	50% 0.555600	50% 0.636400	50% 0.736800	50% 0.727300
			75% 0.857100	75% 0.750000	75% 0.800000	75% 0.857100	75% 0.857100
			max 1.000000	max 0.952400	max 1.000000	max 1.000000	max 1.000000
		N	count 21.000000	count 127.000000	count 3.000000	count 160.000000	count 311.000000
			mean 0.692305	mean 0.658554	mean 0.846167	mean 0.595713	mean 0.630313
			std 0.249470	std 0.228517	std 0.153850	std 0.206086	std 0.220981
			min 0.071400	min 0.076900	min 0.692300	min 0.105300	min 0.071400
			25% 0.588200	25% 0.500000	25% 0.769250	25% 0.473700	25% 0.500000
			50% 0.714300	50% 0.666700	50% 0.846200	50% 0.631600	50% 0.666700
			75% 0.857100	75% 0.839750	75% 0.923100	75% 0.740100	75% 0.789500
			max 1.000000	max 1.000000	max 1.000000	max 1.000000	max 1.000000
	Jaccard	**Y**	count 562.000000	count 79.000000	count 119.000000	count 4119.000000	count 4879.000000
			mean 0.639937	mean 0.516247	mean 0.566463	mean 0.659514	mean 0.652670
			std 0.207119	std 0.217078	std 0.234924	std 0.194258	std 0.198544
			min 0.000000	min 0.071400	min 0.000000	min 0.000000	min 0.000000
			25% 0.500000	25% 0.324550	25% 0.416700	25% 0.538500	25% 0.533300
			50% 0.666700	50% 0.526300	50% 0.583300	50% 0.687500	50% 0.666700
			75% 0.800000	75% 0.692300	75% 0.727300	75% 0.800000	75% 0.800000
			max 1.000000	max 0.909100	max 1.000000	max 1.000000	max 1.000000
		N	count 21.000000	count 127.000000	count 3.000000	count 160.000000	count 311.000000
			mean 0.652705	mean 0.618998	mean 0.809533	mean 0.562634	mean 0.594114
			std 0.245617	std 0.223707	std 0.179739	std 0.194924	std 0.213017
			min 0.066700	min 0.071400	min 0.642900	min 0.100000	min 0.066700
			25% 0.555600	25% 0.461500	25% 0.714300	25% 0.453375	25% 0.464100
			50% 0.666700	50% 0.615400	50% 0.785700	50% 0.600000	50% 0.615400
			75% 0.800000	75% 0.789500	75% 0.892850	75% 0.700000	75% 0.750000
			max 1.000000	max 1.000000	max 1.000000	max 1.000000	max 1.000000

Smooth	**Hamming**	**Y**	count 562.000000	count 79.000000	count 119.000000	count 4119.000000	count 4879.000000
			mean 0.678622	mean 0.550406	mean 0.614651	mean 0.706830	mean 0.923064
			std 0.214634	std 0.229354	std 0.243665	std 0.199468	std 0.137383
			min 0.000000	min 0.076900	min 0.000000	min 0.000000	min 0.055600
			25% 0.538500	25% 0.348450	25% 0.454500	25% 0.583300	25% 0.909100
			50% 0.699100	50% 0.555600	50% 0.636400	50% 0.736800	50% 1.000000
			75% 0.857100	75% 0.750000	75% 0.800000	75% 0.857100	75% 1.000000
			max 1.000000	max 0.952400	max 1.000000	max 1.000000	max 1.000000
		N	count 21.000000	count 127.000000	count 3.000000	count 160.000000	count 311.000000
			mean 0.692305	mean 0.658554	mean 0.846167	mean 0.595713	mean 0.889486
			std 0.249470	std 0.228517	std 0.153850	std 0.206086	std 0.159640
			min 0.071400	min 0.076900	min 0.692300	min 0.105300	min 0.214300
			25% 0.588200	25% 0.500000	25% 0.769250	25% 0.473700	25% 0.846200
			50% 0.714300	50% 0.666700	50% 0.846200	50% 0.631600	50% 0.944400
			75% 0.857100	75% 0.839750	75% 0.923100	75% 0.740100	75% 1.000000
			max 1.000000	max 1.000000	max 1.000000	max 1.000000	max 1.000000
	Jaccard	**Y**	count 562.000000	count 79.000000	count 119.000000	count 4119.000000	count 4879.000000
			mean 0.639937	mean 0.516247	mean 0.566463	mean 0.659514	mean 0.897143
			std 0.207119	std 0.217078	std 0.234924	std 0.194258	std 0.154479
			min 0.000000	min 0.071400	min 0.000000	min 0.000000	min 0.052600
			25% 0.500000	25% 0.324550	25% 0.416700	25% 0.538500	25% 0.833300
			50% 0.666700	50% 0.526300	50% 0.583300	50% 0.687500	50% 1.000000
			75% 0.800000	75% 0.692300	75% 0.727300	75% 0.800000	75% 1.000000
			max 1.000000	max 0.909100	max 1.000000	max 1.000000	max 1.000000
		N	count 21.000000	count 127.000000	count 3.000000	count 160.000000	count 311.000000
			mean 0.652705	mean 0.618998	mean 0.809533	mean 0.562634	mean 0.861658
			std 0.245617	std 0.223707	std 0.179739	std 0.194924	std 0.170498
			min 0.066700	min 0.071400	min 0.642900	min 0.100000	min 0.200000
			25% 0.555600	25% 0.461500	25% 0.714300	25% 0.453375	25% 0.794750
			50% 0.666700	50% 0.615400	50% 0.785700	50% 0.600000	50% 0.894700
			75% 0.800000	75% 0.789500	75% 0.892850	75% 0.700000	75% 1.000000
			max 1.000000	max 1.000000	max 1.000000	max 1.000000	max 1.000000
GradCAM	**Hamming**	**Y**	count 562.000000	count 79.000000	count 119.000000	count 4119.000000	count 4879.000000
			mean 0.678622	mean 0.550406	mean 0.614651	mean 0.706830	mean 0.698799
			std 0.214634	std 0.229354	std 0.243665	std 0.199468	std 0.204438
			min 0.000000	min 0.076900	min 0.000000	min 0.000000	min 0.000000
			25% 0.538500	25% 0.348450	25% 0.454500	25% 0.583300	25% 0.571400
			50% 0.699100	50% 0.555600	50% 0.636400	50% 0.736800	50% 0.727300
			75% 0.857100	75% 0.750000	75% 0.800000	75% 0.857100	75% 0.857100
			max 1.000000	max 0.952400	max 1.000000	max 1.000000	max 1.000000
		N	count 21.000000	count 127.000000	count 3.000000	count 160.000000	count 311.000000
			mean 0.692305	mean 0.658554	mean 0.846167	mean 0.595713	mean 0.630313
			std 0.249470	std 0.228517	std 0.153850	std 0.206086	std 0.220981
			min 0.071400	min 0.076900	min 0.692300	min 0.105300	min 0.071400
			25% 0.588200	25% 0.500000	25% 0.769250	25% 0.473700	25% 0.500000
			50% 0.714300	50% 0.666700	50% 0.846200	50% 0.631600	50% 0.666700
			75% 0.857100	75% 0.839750	75% 0.923100	75% 0.740100	75% 0.789500
			max 1.000000	max 1.000000	max 1.000000	max 1.000000	max 1.000000

	Jaccard	Y	count 562.000000	count 79.000000	count 119.000000	count 4119.000000	count 4879.000000
			mean 0.639937	mean 0.516247	mean 0.566463	mean 0.659514	mean 0.652670
			std 0.207119	std 0.217078	std 0.234924	std 0.194258	std 0.198544
			min 0.000000	min 0.071400	min 0.000000	min 0.000000	min 0.000000
			25% 0.500000	25% 0.324550	25% 0.416700	25% 0.538500	25% 0.533300
			50% 0.666700	50% 0.526300	50% 0.583300	50% 0.687500	50% 0.666700
			75% 0.800000	75% 0.692300	75% 0.727300	75% 0.800000	75% 0.800000
			max 1.000000	max 0.909100	max 1.000000	max 1.000000	max 1.000000
		N	count 21.000000	count 127.000000	count 3.000000	count 160.000000	count 311.000000
			mean 0.652705	mean 0.618998	mean 0.809533	mean 0.562634	mean 0.594114
			std 0.245617	std 0.223707	std 0.179739	std 0.194924	std 0.213017
			min 0.066700	min 0.071400	min 0.642900	min 0.100000	min 0.066700
			25% 0.555600	25% 0.461500	25% 0.714300	25% 0.453375	25% 0.464100
			50% 0.666700	50% 0.615400	50% 0.785700	50% 0.600000	50% 0.615400
			75% 0.800000	75% 0.789500	75% 0.892850	75% 0.700000	75% 0.750000
			max 1.000000	max 1.000000	max 1.000000	max 1.000000	max 1.000000
Grad-CAM++	Hamming	Y	count 562.000000	count 79.000000	count 119.000000	count 4119.000000	count 4879.000000
			mean 0.678622	mean 0.550406	mean 0.614651	mean 0.706830	mean 0.698799
			std 0.214634	std 0.229354	std 0.243665	std 0.199468	std 0.204438
			min 0.000000	min 0.076900	min 0.000000	min 0.000000	min 0.000000
			25% 0.538500	25% 0.348450	25% 0.454500	25% 0.583300	25% 0.571400
			50% 0.699100	50% 0.555600	50% 0.636400	50% 0.736800	50% 0.727300
			75% 0.857100	75% 0.750000	75% 0.800000	75% 0.857100	75% 0.857100
			max 1.000000	max 0.952400	max 1.000000	max 1.000000	max 1.000000
		N	count 21.000000	count 127.000000	count 3.000000	count 160.000000	count 311.000000
			mean 0.692305	mean 0.658554	mean 0.846167	mean 0.595713	mean 0.630313
			std 0.249470	std 0.228517	std 0.153850	std 0.206086	std 0.220981
			min 0.071400	min 0.076900	min 0.692300	min 0.105300	min 0.071400
			25% 0.588200	25% 0.500000	25% 0.769250	25% 0.473700	25% 0.500000
			50% 0.714300	50% 0.666700	50% 0.846200	50% 0.631600	50% 0.666700
			75% 0.857100	75% 0.839750	75% 0.923100	75% 0.740100	75% 0.789500
			max 1.000000	max 1.000000	max 1.000000	max 1.000000	max 1.000000
	Jaccard	Y	count 562.000000	count 79.000000	count 119.000000	count 4119.000000	count 4879.000000
			mean 0.639937	mean 0.516247	mean 0.566463	mean 0.659514	mean 0.652670
			std 0.207119	std 0.217078	std 0.234924	std 0.194258	std 0.198544
			min 0.000000	min 0.071400	min 0.000000	min 0.000000	min 0.000000
			25% 0.500000	25% 0.324550	25% 0.416700	25% 0.538500	25% 0.533300
			50% 0.666700	50% 0.526300	50% 0.583300	50% 0.687500	50% 0.666700
			75% 0.800000	75% 0.692300	75% 0.727300	75% 0.800000	75% 0.800000
			max 1.000000	max 0.909100	max 1.000000	max 1.000000	max 1.000000
		N	count 21.000000	count 127.000000	count 3.000000	count 160.000000	count 311.000000
			mean 0.652705	mean 0.618998	mean 0.809533	mean 0.562634	mean 0.594114
			std 0.245617	std 0.223707	std 0.179739	std 0.194924	std 0.213017
			min 0.066700	min 0.071400	min 0.642900	min 0.100000	min 0.066700
			25% 0.555600	25% 0.461500	25% 0.714300	25% 0.453375	25% 0.464100
			50% 0.666700	50% 0.615400	50% 0.785700	50% 0.600000	50% 0.615400
			75% 0.800000	75% 0.789500	75% 0.892850	75% 0.700000	75% 0.750000
			max 1.000000	max 1.000000	max 1.000000	max 1.000000	max 1.000000

Table A.4: Novelty - Vanilla Saliency - Full Results

SR			**GSVT**			**AFIB**			**SB**		
TF-IDF	**Feature**	**Record**	**TF-IDF**	**Feature**	**Record**	**TF-IDF**	**Feature**	**Record**	**TF-IDF**	**Feature**	**Record**
0.00000	0	0	0.02404	0	0	0.02947	0	0			
0.41448	13	0	0.00000	1	0	0.00000	1	0			
0.44928	37	0	0.49167	4	0	0.29287	44	0			
0.39364	112	0	0.41003	66	0	0.25937	83	0			
0.41954	125	0	0.36227	240	0	0.27090	105	0			
0.41783	170	0	0.05129	311	0	0.25426	181	0			
0.42669	205	0	0.46485	358	0	0.23692	213	0			
0.40355	262	0	0.55741	393	0	0.22629	277	0			
0.46667	447	0	0.41071	206	1	0.23319	321	0			
0.43841	492	0	0.37230	373	1	0.28474	391	0			
0.01573	0	1	0.36679	245	2	0.04501	496	0			
0.34367	25	1	0.43982	275	2	0.44725	57	1			
0.34506	71	1	0.29375	405	2	0.38102	75	1			
0.36037	117	1	0.32407	438	2	0.36783	182	1			
0.33962	161	1	0.03146	55	3	0.42374	217	1			
0.33831	249	1	0.42979	123	3	0.39609	263	1			
0.35708	294	1	0.53664	174	3	0.33598	360	1			
0.36733	343	1	0.64348	217	3	0.06585	492	1			
0.33575	388	1	0.05871	366	3	0.55746	84	2			
0.34792	477	1	0.30936	198	4	0.52935	165	2			
0.02414	0	2	0.22794	213	4	0.61996	282	2			
0.54314	97	2	0.25799	319	4	0.60527	408	2			
0.50236	143	2	0.02992	488	4	0.03166	488	2			
0.49221	191	2	0.21412	493	4	0.31458	32	3			
0.55563	462	2	0.17856	496	4	0.27311	57	3			
0.01723	0	3	0.13156	8	5	0.28453	82	3			
0.38433	31	3	0.15777	118	5	0.30585	101	3			
0.37952	77	3	0.01513	299	5	0.33590	119	3			
0.36241	122	3	0.17657	384	5	0.25872	230	3			
0.34352	166	3	0.20662	435	5	0.25048	348	3			
0.35368	223	3	0.04238	421	6	0.23961	383	3			
0.41509	300	3	0.27865	235	7	0.25449	430	3			
0.35857	393	3	0.33413	249	7	0.12290	484	3			
0.53617	105	4	0.19071	317	7	1.42059	245	4			
0.53172	152	4	0.22317	365	7	1.08151	365	4			
0.49719	197	4	0.24619	433	7	0.37422	102	5			
0.50592	242	4	0.01634	354	8	0.43484	214	5			
0.56372	285	4	0.02270	165	9	0.35071	368	5			
0.52531	374	4	0.19898	174	9	0.03419	476	5			
0.35087	3	5	0.15936	192	9	0.32191	23	6			
0.38105	45	5	0.23860	224	9	0.28946	53	6			
0.36554	90	5	0.13619	250	9	0.33035	90	6			
0.32621	132	5	0.01237	476	9	0.30091	124	6			
0.33205	176	5	0.18647	478	9	0.31428	190	6			
0.31167	225	5	0.21105	492	9	0.37740	226	6			
0.38540	402	5	0.25307	50	10	0.03717	472	6			

0.39659	11	6
0.37645	104	6
0.36505	151	6
0.37201	196	6
0.41285	241	6
0.37347	328	6
0.36641	373	6
0.39292	419	6
0.40642	466	6
0.02787	0	7
0.64759	90	7
0.58614	136	7
0.55217	180	7
0.61637	317	7
0.59483	362	7
0.02129	0	8
0.47700	102	8
0.44311	149	8
0.43855	194	8
0.45796	238	8
0.45280	280	8
0.43561	369	8
0.01340	0	9
0.32637	41	9
0.32101	57	9
0.28705	84	9
0.27227	129	9
0.27500	173	9
0.29389	218	9
0.29154	265	9
0.28490	310	9
0.29039	354	9
0.31764	396	9
0.29757	452	9
0.30413	496	9
0.38599	20	10
0.35488	194	10
0.38939	237	10
0.05201	0	11
1.13626	35	11
1.03687	171	11
0.01447	0	12
0.32005	14	12
0.33306	61	12
0.31242	106	12
0.30113	149	12
0.33151	190	12
0.28243	193	12
0.31122	238	12
0.31364	328	12
0.16903	141	10
0.02921	287	10
0.30291	131	11
0.26763	149	11
0.36322	207	11
0.32839	42	12
0.24260	288	12
0.47413	225	13
1.20090	164	14
0.88485	176	14
1.00149	371	14
0.02725	275	15
0.20367	410	16
0.15007	420	16
0.16985	439	16
0.13603	148	17
0.01776	330	18
0.09593	452	19
0.39786	14	20
0.33180	79	20
0.04554	263	20
0.07505	74	21
0.19008	405	21
0.09909	196	22
0.06867	7	23
0.26573	75	23
0.29315	262	23
0.06330	440	24
0.04096	251	25
0.27024	271	25
0.03409	62	26
0.12634	495	27
0.31816	80	28
0.11421	306	28
0.08275	117	29
0.08885	428	29
0.20531	381	31
0.12371	172	32
0.07647	227	35
0.22318	247	35
0.12830	38	36
0.09018	349	36
0.20732	489	39
0.32066	41	7
0.31466	78	7
0.32711	112	7
0.30377	150	7
0.34994	224	7
0.33408	288	7
0.34165	369	7
0.38097	465	7
0.56809	93	8
0.59185	128	8
0.65436	167	8
0.69870	246	8
0.52102	330	8
0.02757	464	8
0.24779	492	8
0.25341	26	9
0.26636	49	9
0.22531	114	9
0.21482	202	9
0.23784	233	9
0.27396	311	9
0.25956	413	9
0.30732	2	10
0.26643	198	10
0.45074	397	10
0.36936	482	10
0.29409	146	11
0.29880	327	11
0.41369	126	12
0.38829	222	12
0.40453	338	12
0.57952	32	13
0.22923	6	15
0.22162	8	15
0.24263	159	15
0.23339	352	15
0.05032	436	15
0.42594	489	15
0.41834	35	16
0.41120	122	16
0.46247	285	16
0.04072	432	16
0.36594	471	16
0.40458	499	16
0.34467	69	17
0.30817	125	17
0.33274	161	17
0.38332	225	17
0.47065	265	17
0.41731	342	17

0.28584	375	12
0.32702	465	12
0.03296	0	13
0.67886	140	13
0.66551	186	13
0.73208	230	13
0.75510	405	13
0.35238	74	14
0.30978	334	14
0.33450	344	14
0.43135	430	14
0.01905	0	15
0.41124	44	15
0.44049	92	15
0.39777	139	15
0.44701	232	15
0.41614	277	15
0.42306	364	15
0.43239	496	15
0.34230	13	16
0.33702	84	16
0.31968	129	16
0.32288	173	16
0.33085	219	16
0.32735	266	16
0.36916	313	16
0.36206	400	16
0.54078	74	17
0.49385	120	17
0.53845	163	17
0.50958	206	17
0.54554	252	17
0.51721	392	17
0.52120	106	18
0.55303	287	18
0.35549	5	19
0.37690	311	19
0.35392	356	19
0.67656	155	20
0.68353	198	20
0.69078	242	20
0.72901	476	20
0.60885	113	21
0.57397	140	21
0.56268	186	21
0.68678	274	21
0.60168	106	22
0.59042	151	22
0.54392	193	22
1.13173	13	23
0.29098	90	18
0.27860	131	18
0.29805	339	18
0.24668	391	18
0.35125	472	18
0.29498	78	20
0.27944	98	20
0.41025	166	21
0.35911	371	21
0.26321	76	22
0.26800	185	22
0.41732	345	22
0.48611	464	22
0.44279	50	23
0.40447	88	23
0.47370	160	23
0.28429	291	24
0.44938	417	24
0.37423	195	25
0.32214	260	25
0.33852	287	25
0.29253	107	26
0.07789	392	26
0.67191	168	27
0.62609	263	27
0.33977	444	28
0.27487	137	29
0.50622	324	29
0.36181	109	31
0.34066	271	31
0.24306	141	34
0.05705	360	34
0.56683	398	34
0.48289	440	34
0.51269	31	35
0.50200	194	35
0.52431	257	35
0.46132	171	36
0.35611	218	36
0.49677	0	38
0.28258	348	38
0.02191	340	39
0.20629	343	39
0.19283	364	39
0.22459	373	39
0.19694	411	39
0.18903	442	39
0.18549	496	39
0.17907	72	40
0.24184	123	40

1.10595	373	23
0.67888	68	24
0.57792	229	24
0.67507	367	24
0.61382	410	24
0.49010	40	25
0.42855	158	25
0.44786	359	25
0.46153	461	25
0.62701	86	26
0.64448	167	26
0.58827	206	26
0.57593	377	26
0.69512	34	27
0.65077	46	27
0.68278	447	27
0.07331	0	28
1.51488	88	28
0.32416	20	29
0.30008	62	29
0.31615	104	29
0.30219	145	29
0.29803	184	29
0.31743	226	29
0.33960	313	29
0.28497	334	29
0.31489	335	29
0.52475	9	30
0.45449	51	30
0.47294	94	30
0.44467	139	30
0.47096	350	30
0.50225	466	30
0.47908	178	31
0.42190	225	31
0.43416	353	31
0.48120	395	31
0.01167	0	32
0.27662	19	32
0.25594	24	32
0.25389	58	32
0.26251	97	32
0.23869	138	32
0.24540	177	32
0.24030	216	32
0.24196	254	32
0.25914	293	32
0.26025	330	32
0.26606	401	32
0.27113	409	32
0.25329	269	40
0.21169	302	40
0.44448	394	40
0.55110	192	41
0.53703	220	41
0.47427	321	41
0.67494	56	42
0.17614	396	42
0.21774	436	42
0.20140	123	43
0.18218	207	43
0.24950	416	43
0.24505	488	43
0.26489	15	44
0.27747	162	44
0.31272	256	44
0.01899	320	44
0.15024	322	44
0.17877	323	44
0.20148	350	44
0.18346	376	44
0.16075	400	44
0.14796	424	44
0.18869	448	44
0.16382	473	44
0.15519	498	44
0.17067	59	45
0.16711	66	45
0.17454	144	45
0.02310	316	45
0.18876	317	45
0.22314	350	45
0.20759	402	45
0.18274	426	45
0.19552	451	45
0.21229	478	45
0.21745	147	46
0.23673	226	46
0.24086	54	47
0.30208	307	47
0.21951	335	47
0.15788	407	47
0.39526	314	48
0.25525	168	49
0.02590	300	49
0.25735	308	49
0.21532	354	49
0.21924	379	49
0.27479	403	49
0.23277	430	49

0.26367	479	32
0.31442	60	33
0.30277	100	33
0.26305	142	33
0.27593	184	33
0.29271	228	33
0.27317	353	33
0.31133	368	33
0.28179	445	33
0.46400	41	34
0.40966	164	34
0.40208	206	34
0.39919	382	34
0.37797	91	35
0.34140	180	35
0.38110	227	35
0.37058	397	35
0.57314	442	35
0.38270	457	35
0.32850	145	36
0.34938	324	36
0.37295	407	36
0.43435	141	37
0.37740	225	37
0.44480	235	37
0.39638	266	37
0.38837	353	37
0.43045	395	37
0.42486	468	37
0.50774	136	38
0.47831	180	38
0.53393	350	38
0.35017	144	39
0.41974	473	39
0.52790	28	40
0.42719	159	40
0.43707	223	40
0.49243	357	40
0.44005	377	40
0.40062	122	41
0.39500	165	41
0.45396	491	41
0.64142	130	42
0.61132	218	42
0.59708	392	42
0.04034	0	43
0.86105	103	43
0.86431	231	43
0.95153	466	43
0.72600	109	44
0.23804	474	49
0.25021	16	50
0.19332	164	50
0.31681	236	50
0.22791	250	50
0.24382	278	50
0.25880	477	53
0.35046	142	54
0.36280	268	54
0.49998	343	55
0.45222	370	55
0.39810	136	56
0.39206	239	56
0.28963	10	57
0.30429	279	58
0.21814	454	58
0.31870	287	59
0.46618	257	60
0.49210	246	61
0.20580	112	63
0.27054	61	65
0.20491	238	65
0.22342	389	65
0.19602	429	65
0.21165	465	65
0.28584	41	66
0.17322	232	66
1.52431	105	67
0.20180	116	68
0.02084	224	68
0.20711	236	68
0.20136	281	68
0.17033	287	68
0.19622	311	68
0.16491	334	68
0.16240	360	68
0.18733	384	68
0.17981	408	68
0.24093	485	68
0.18342	35	69
0.19157	87	69
0.17329	111	69
0.26805	159	70
0.22965	444	72
0.22001	68	74
0.39439	441	74
0.30904	185	75
0.26253	469	76
0.32530	336	78
0.32448	205	83

0.64911	334	44
0.67430	369	44
0.76226	423	45
0.71725	489	45
0.43636	39	46
0.40810	84	46
0.39230	216	46
0.43523	6	47
0.40237	47	47
0.35609	88	47
0.34245	217	47
0.56822	220	48
0.60642	459	48
0.71327	8	49
0.48781	26	50
0.44948	243	50
0.51019	289	50
0.57993	131	51
0.56088	175	51
0.55048	375	51
0.34095	147	52
0.58155	21	53
0.48900	195	53
0.41285	253	54
0.47495	3	55
0.46521	228	55
0.46335	459	55
0.57534	27	56
0.49889	112	56
0.48742	199	56
0.55827	245	56
0.52955	483	56
0.40656	255	57
0.44262	494	57
0.43133	195	58
0.49724	235	58
0.46901	371	58
0.52741	35	59
0.48128	171	59
0.54800	263	59
0.34903	195	60
0.39114	327	60
0.37495	374	60
0.39236	79	61
0.32074	220	61
0.37103	394	61
0.50413	160	62
0.52324	253	62
0.71164	161	63
0.74488	252	63
0.43407	337	85
0.39336	57	91
0.28538	409	91
0.45470	68	92
0.31725	260	93
0.32729	304	93
0.35831	338	94
0.53386	211	97
0.09533	92	101
0.47756	111	102
0.19203	158	105
0.20326	233	105
0.26699	445	105
0.22951	1	106
0.34941	23	107
0.06113	68	107
0.36595	329	107
0.51311	122	108
0.54751	443	108
0.49842	459	108
0.30393	424	109
0.15709	56	110
0.17733	313	115
0.24507	20	116
0.05347	28	117
0.61623	26	118
0.75247	121	118
0.71588	337	118
0.73326	431	118
0.08573	20	119
0.10736	16	120
0.07137	8	122
0.23632	364	124
0.34557	87	126
0.02442	484	127
0.25908	3	128
0.20670	20	128
0.23590	40	128
0.21946	113	128
0.21488	132	128
0.19628	197	128
0.19026	220	128
0.19319	292	128
0.19955	328	128
0.21065	436	128
0.20301	443	128
0.24263	83	130
0.22988	228	130
0.20818	205	131
0.76627	161	137

0.70091	344	63
0.72304	483	63
0.38710	144	64
0.48427	168	64
0.39097	188	64
0.42855	264	64
0.72012	35	65
0.65307	180	65
0.42452	166	66
0.50484	396	66
0.53112	488	66
0.38459	199	67
1.60147	113	68
0.43247	18	69
0.38767	63	69
0.35732	198	69
0.36372	243	69
0.40437	329	69
0.41808	142	71
0.42993	186	71
0.43274	233	71
0.47511	34	72
0.37974	166	72
0.42128	211	72
0.37398	142	73
0.35250	138	74
0.36778	182	74
0.39474	190	74
0.38998	28	75
0.34648	70	75
0.31659	158	75
0.39848	22	76
0.35983	160	76
0.49551	188	77
0.51334	373	77
0.42320	217	78
0.45113	262	78
0.49481	494	78
0.32509	112	80
0.42463	307	83
0.32398	140	84
0.31072	375	84
0.40504	310	85
0.37490	27	86
0.46709	91	89
0.49972	313	89
0.58404	244	90
0.59936	291	90
0.46941	9	93
0.59492	307	94
0.97702	287	137
0.85675	354	137
0.28134	137	138
0.41828	21	141
0.43175	287	151
0.34069	411	152
0.17074	452	161
0.23249	274	162
0.38727	267	167
0.76121	347	168
0.50559	97	169
0.43200	49	172
0.31260	81	173
0.65933	7	174
0.70001	178	174
0.94723	392	177
0.89743	86	178
0.77903	128	178
0.62959	215	193
0.45133	68	197
0.82235	451	210
0.63620	153	215
0.45855	361	221
0.77394	127	229
0.26544	57	245
0.53816	208	247
0.21823	0	248
0.91738	364	248
0.40816	415	252
0.14374	420	267
0.68542	257	268
0.22308	101	271
0.29937	331	274
0.49157	236	280
0.80696	453	286
0.92094	123	287
0.63651	309	295
0.64756	267	296
0.13605	260	307
0.58162	338	307
0.51571	381	307
0.58466	494	312
0.60524	224	313
0.11208	236	313
0.11460	232	314
0.42967	93	351
0.37559	328	381
0.36054	351	394
0.52028	19	395
1.49426	385	417

0.37625	375	95
0.67135	21	97
0.61896	297	97
0.59261	475	97
0.51526	182	98
0.67411	431	98
0.78556	491	99
0.45973	372	101
0.41934	334	103
0.48337	345	103
0.54882	334	105
0.57203	155	106
0.62428	381	106
0.44625	383	107
0.35133	213	109
0.59145	17	110
0.50062	198	110
0.56653	329	110
0.48557	401	112
0.32276	111	113
0.32140	154	113
0.29312	195	113
0.30435	239	113
0.34132	282	113
0.30886	485	113
0.38320	23	115
0.32967	242	115
0.34036	375	116
0.49060	258	117
0.53444	18	119
0.51580	367	119
0.51871	473	119
0.28671	225	122
0.32848	263	122
0.31873	305	122
0.62979	100	123
0.58197	145	123
0.38214	159	124
0.92851	42	125
0.80933	159	125
0.86762	463	125
0.56941	389	128
0.36378	22	129
0.35871	73	129
0.34308	16	130
0.30546	114	130
0.35875	157	130
0.32557	264	130
0.33626	409	130
0.51919	164	132
1.55676	410	417
0.79829	353	419
0.55059	329	436
0.83889	383	439
0.74276	464	439
0.79255	431	444
0.25111	192	448
0.22443	204	449
0.25027	412	449
0.75421	28	455
0.49807	241	458
0.38083	499	470
0.60688	230	498
0.27747	332	504
0.72681	399	509
1.02178	167	516
1.18730	406	516
0.16823	495	520
0.36149	310	528
0.20867	215	541

0.59848	411	132
0.06448	0	133
0.44157	165	134
0.45621	210	134
0.58389	429	137
0.45638	57	139
0.32454	45	140
0.29632	93	140
0.27408	138	140
0.27051	186	140
0.31286	285	140
0.41559	436	140
0.40823	279	141
0.32180	138	143
0.47672	426	144
0.59086	433	146
1.25984	124	148
1.07482	380	148
0.51297	300	149
0.36790	56	150
0.29505	220	150
0.30771	275	150
0.39680	429	150
0.52170	402	152
0.65733	389	153
0.34790	162	156
0.40852	333	156
0.40040	494	156
0.56635	144	160
0.47398	142	161
0.58477	367	161
0.91575	183	162
0.89981	468	162
0.56097	409	163
0.65713	166	164
0.65509	217	164
0.68833	139	165
0.67206	247	165
0.60214	488	166
0.42979	146	167
0.33630	193	167
0.62160	209	170
0.54717	142	171
0.56451	195	171
0.45886	89	172
0.73837	20	174
0.68592	131	174
0.76594	250	174
0.51145	276	176
0.53786	6	177

0.69946	18	178
0.63261	352	178
0.63843	26	180
0.57011	138	180
0.42717	157	181
0.60403	319	182
0.43808	56	188
0.34679	175	188
0.36917	288	188
1.59509	13	189
1.53622	244	189
0.82254	353	192
0.94203	458	192
0.47222	28	195
0.70353	342	196
0.29408	144	197
0.31003	288	197
0.34488	339	197
0.66196	430	201
0.41066	487	204
1.12288	361	205
1.19147	471	205
0.44101	8	209
0.39993	56	210
0.31361	166	210
0.52324	442	210
0.38335	158	211
0.39481	425	212
0.33327	151	213
0.43926	431	214
0.46140	484	215
0.46292	438	217
0.47117	193	218
0.61787	482	218
0.37895	300	219
0.42216	402	226
0.45160	396	228
0.41739	367	233
0.38967	120	238
0.37511	334	238
0.48585	99	245
1.16005	163	250
1.10187	276	250
0.89599	105	251
0.83370	112	251
0.89224	156	251
1.20275	490	253
0.35053	124	255
0.30658	187	255
0.34672	415	255

1.04673	175	256
1.09388	219	256
0.33017	9	257
0.27978	139	257
0.34063	168	257
0.28926	185	257
0.28596	406	257
0.30837	462	257
0.30551	141	259
0.33215	157	259
0.27783	229	259
0.29883	314	259
0.55910	159	260
0.88492	91	264
0.82527	138	264
0.97721	367	264
0.47685	375	265
0.39734	6	266
0.35251	41	269
0.29905	88	269
0.31761	199	272
0.48752	482	277
0.28814	75	279
0.30983	256	279
0.32275	300	279
0.27139	456	279
0.36551	6	283
0.30326	242	283
0.30702	193	285
0.29702	155	286
0.31864	195	289
0.50749	487	291
0.83658	62	292
0.82805	155	292
0.48113	6	294
0.30143	97	299
0.30012	325	299
0.33627	418	299
0.38766	72	302
0.40261	8	306
0.26545	225	309
0.48431	159	310
0.70890	249	312
0.74159	295	312
0.43521	427	314
0.34568	159	315
0.33464	92	320
0.32998	401	320
0.32825	96	322
0.27880	143	322

0.36738	432	322
0.60977	6	330
0.04544	0	332
0.58807	41	332
0.55260	279	337
0.57234	19	344
0.38583	195	351
0.29124	175	352
0.48099	437	354
0.66418	339	356
0.31264	217	357
0.28849	166	359
0.36111	121	363
0.55049	401	365
0.47807	18	366
0.81715	157	372
0.77746	261	372
0.35453	96	378
0.33791	341	378
1.15513	457	382
0.04850	0	391
0.28281	181	391
0.25289	29	393
0.23558	162	393
0.24629	206	393
0.26982	245	393
0.27246	341	393
0.07870	0	396
0.03754	0	397
0.29510	70	397
0.31601	346	397
0.27688	377	397
0.08172	0	398
0.28063	82	398
0.24867	166	398
0.26329	208	398
0.28187	252	398
0.27941	412	398
0.25265	456	398
0.37176	193	399
0.29603	138	403
0.36094	79	406
0.26729	26	408
0.25806	65	408
0.25491	104	408
0.23712	144	408
0.26486	183	408
0.24719	262	408
0.24811	386	408
0.26137	412	408

0.29284	425	408
0.28263	484	408
0.63550	419	410
0.69089	413	416
0.01247	0	420
0.29572	19	420
0.27361	137	420
0.27035	215	420
0.25602	223	420
0.26623	255	420
0.29420	296	420
0.27142	335	420
0.25777	380	420
0.28443	419	420
0.29217	186	421
0.35661	390	421
0.81452	162	422
0.96658	241	422
0.26855	61	429
0.22908	142	429
0.27521	261	429
0.24452	340	429
0.25094	379	429
0.33418	79	438
0.28385	151	438
0.26149	193	440
0.31931	487	440
0.66072	333	441
0.66773	289	443
0.27587	14	444
0.25689	140	444
0.24713	180	444
0.26234	219	444
0.25348	258	444
0.27703	297	444
0.24563	334	444
0.29727	339	444
0.30384	473	444
0.30214	484	444
0.30885	142	445
0.30693	26	446
0.27821	31	448
0.24344	193	448
0.26723	231	448
0.28574	270	448
0.34202	430	448
0.28845	490	448
0.02263	0	451
0.29471	6	451
0.25699	77	451

0.23949	155	451
0.23634	195	451
0.23790	247	451
0.29272	60	452
0.26826	249	452
0.28314	327	452
0.30559	17	453
0.26929	196	453
0.26425	276	453
0.25956	393	453
0.34830	431	453
0.28708	498	453
0.37856	217	457
1.57651	397	458
0.30922	30	461
0.33356	101	461
0.26523	310	461
0.24998	288	464
0.28108	300	464
0.27807	339	464
0.24113	377	464
0.25432	247	465
0.26140	383	466
0.27508	18	467
0.24042	24	467
0.23850	58	467
0.24659	97	467
0.22421	138	467
0.23135	176	467
0.23052	214	467
0.23393	251	467
0.24768	272	467
0.23482	288	467
0.24343	293	467
0.26549	367	467
0.25108	414	467
0.24241	477	467
0.34860	322	468
0.36319	83	469
0.03510	0	471
0.27250	25	471
0.25866	62	471
0.25103	99	471
0.25516	120	478
0.27473	371	478
0.37067	426	479
0.24904	103	481
0.30068	318	481
0.24280	393	481
0.25191	454	481

0.01096	0	482
0.24552	20	482
0.23945	113	482
0.22274	129	482
0.22728	165	482
0.23306	203	482
0.24446	314	482
0.22888	384	482
0.25469	450	482
0.31711	56	484
0.31924	416	484
0.83801	168	486
0.25721	15	487
0.26260	57	487
0.21988	159	487
0.23663	196	487
0.22573	271	487
0.27010	307	487
0.22808	376	487
0.24141	410	487
0.25346	423	487
0.22497	1	488
0.27684	6	488
0.25226	42	488
0.22347	220	488
0.25985	333	488
0.36193	433	490
0.26627	217	491
0.24638	375	493
0.30048	499	493
0.28754	9	494
0.23482	158	494
0.27382	417	494
0.28078	244	497
0.27956	241	499
0.03020	0	501
0.28586	17	506
0.28984	368	510
0.03628	0	511
0.33933	334	516
0.47978	217	517
0.69326	359	527
0.52699	426	530
1.27423	17	534
1.07853	132	534
0.87097	204	535
0.87438	319	535
0.85784	373	535
0.75163	134	538
0.26710	166	560

0.48133	442	571
0.74528	101	574
0.47541	334	576
0.55559	171	581
0.37037	8	584
0.38699	101	584
0.51299	441	591
0.04743	0	594
0.44716	279	595
0.06083	0	597
0.79844	45	600
0.73519	468	600
0.84362	482	600
0.81191	99	605
0.84245	160	605
0.78146	464	610
0.57841	311	613
0.76970	458	627
1.14087	126	629
1.11009	172	629
0.68118	112	634
0.69577	192	634
0.71442	76	644
0.90764	55	653
0.83949	116	653
0.88855	174	653
0.75864	38	654
0.66338	158	654
0.88135	104	660
0.70619	84	663
0.08497	0	676
0.79404	89	681
1.07116	80	682
1.12727	147	682
0.70853	56	683
0.81229	9	685
0.74822	115	685
0.66985	144	688
0.66127	159	691
0.65402	15	697
0.91166	53	705
0.85156	176	705
0.80294	41	714
0.55387	217	717
0.92417	33	720
0.84848	169	720
1.08610	118	726
0.31111	79	727
0.26047	145	727
0.26464	375	730

0.34291	482	730
0.40966	437	733
0.29126	458	734
0.29885	495	734
0.26121	12	735
0.22650	88	735
0.21850	166	735
0.25852	282	735
0.24880	323	735
0.32254	438	735
0.34509	426	736
0.30738	390	743
0.25023	159	745
0.36706	438	745
0.31306	83	748
0.39291	439	748
0.26964	175	749
0.24789	217	751
0.36754	435	754
0.29103	422	757
0.42204	439	759
0.23047	375	762
0.37412	431	767
0.29863	482	769
0.28760	217	771
0.26384	334	772
0.36423	428	776
0.23118	180	779
0.26879	159	788
0.37291	318	793
0.30279	279	794
0.34336	438	794
0.24489	142	796
0.41488	442	798
0.52231	430	803
0.25184	162	804
0.40676	441	804
0.42061	375	808
0.31559	159	809
0.31994	430	816
0.37553	279	823
0.34526	318	829
0.71817	318	858
0.46080	101	867
0.81983	144	871
0.85468	151	879
0.53189	431	889
1.05695	129	901
1.18061	183	901
0.40036	427	906

0.91194	426	912
0.41560	193	916
0.51783	428	918
0.79930	180	952
0.33831	142	956
0.46843	428	966
0.83085	194	971
0.98272	473	975
0.77354	232	980
0.69832	262	980
0.87784	150	991
0.50939	101	993
0.59461	431	1012
0.28412	142	1016
0.54874	318	1078
0.47249	432	1089
0.54500	482	1099
0.44404	318	1131
1.46138	166	1173
0.40537	318	1184
1.21449	67	1214
1.06754	188	1214
0.29031	159	1226
0.43473	440	1227
0.40408	431	1235
1.16505	381	1238
0.60590	418	1246
0.57888	428	1251
0.35829	101	1259
1.20856	90	1297
1.08997	244	1297
0.34769	279	1299
0.32369	279	1300
0.81716	456	1344
1.23306	333	1345
1.18599	141	1372
1.11429	210	1372
0.78975	311	1381
0.56945	101	1383
0.65628	428	1385
0.61376	56	1403
0.75762	428	1428
0.62117	441	1453
0.94673	246	1474
0.96145	491	1474
0.53449	433	1478
0.82728	83	1507
0.80755	96	1508
1.17533	272	1509
1.60796	170	1526

0.60002	439	1578
0.93743	50	1619
0.54137	56	1647
0.42764	428	1683
1.25291	89	1686
0.25594	48	1687
0.22969	121	1687
0.22058	158	1687
0.26852	96	1688
0.38200	440	1688
0.24366	139	1695
0.02587	0	1698
0.26403	89	1715
0.23572	238	1715
0.31506	6	1744
0.44247	437	1771
0.32143	318	1774
0.44888	433	1791
0.28926	157	1799
0.33841	6	1801
0.56055	438	1854
1.73790	290	1943
0.23756	328	1947
0.38691	436	1954
0.09235	0	1956
0.28423	473	1960
0.33909	428	1964
0.22129	162	1966
0.30605	431	1968
0.22201	456	1968
0.01033	0	1972
0.23249	63	1972
0.22951	94	1972
0.23049	128	1972
0.20797	158	1972
0.21210	188	1972
0.20474	225	1972
0.21000	258	1972
0.24499	290	1972
0.22397	328	1972
0.20350	334	1972
0.22310	361	1972
0.21504	393	1972
0.25775	422	1972
0.23784	462	1972
0.24894	493	1972
0.22772	193	1973
0.32281	427	1973
0.42585	438	1983
0.80426	166	1994

0.82215	488	2003
0.64332	193	2005
1.15030	123	2020
1.08229	393	2020
0.06648	0	2029
0.91992	73	2036
1.56459	110	2103
0.21519	142	2171
0.33513	437	2171
0.44564	442	2177
0.23189	217	2184
0.40666	440	2187
0.43691	441	2189
0.48795	433	2214
0.06861	0	2247
0.06774	0	2248
0.05883	0	2254
0.04360	0	2256
0.28052	416	2257
0.44731	443	2285
0.24690	6	2294
0.21355	36	2294
0.21619	65	2294
0.21529	125	2294
0.20341	153	2294
0.20633	181	2294
0.20131	202	2294
0.19930	236	2294
0.23547	300	2294
0.21802	330	2294
0.21103	361	2294
0.21022	391	2294
0.22290	419	2294
0.22605	423	2294
0.23811	473	2294
0.22855	14	2295
0.22576	36	2295
0.21974	69	2295
0.21069	191	2295
0.24130	212	2295
0.22056	251	2295
0.25466	307	2295
0.23352	338	2295
0.22224	370	2295
0.28855	431	2295
0.24012	450	2295
0.07788	0	2310
0.24759	57	2314
0.20537	217	2314
0.22139	387	2314

0.23564	401	2314
0.22667	91	2315
0.23897	167	2315
0.20412	375	2315
0.24374	408	2315
0.34372	442	2315
0.21734	445	2315
0.23148	497	2315
0.48047	443	2331
0.22977	334	2336
0.30323	427	2341
0.30054	429	2357
0.23261	166	2364
0.27338	257	2372
0.24233	28	2375
0.24851	56	2375
0.21710	94	2375
0.20412	160	2375
0.19367	225	2375
0.20942	288	2375
0.22089	301	2375
0.21186	319	2375
0.20863	388	2375
0.24381	422	2375
0.22393	471	2375
0.22714	494	2375
0.00978	0	2377
0.21896	20	2377
0.23296	43	2377
0.22498	61	2377
0.20485	121	2377
0.19672	158	2377
0.19864	221	2377
0.21442	347	2377
0.23420	495	2377
0.64716	142	2384
0.29664	56	2391
0.32581	431	2403
0.28244	318	2412
0.27172	28	2426
0.34525	439	2426
0.21649	375	2433
0.20558	122	2434
0.22188	183	2434
0.24088	274	2434
0.23174	333	2434
0.23055	394	2434
0.20709	262	2435
0.31876	441	2435
0.22825	458	2435

0.51053	444	2436
0.00927	0	2438
0.20099	29	2438
0.22218	57	2438
0.20773	59	2438
0.22339	89	2438
0.18970	120	2438
0.19646	151	2438
0.18373	180	2438
0.19574	208	2438
0.20510	240	2438
0.19099	271	2438
0.20684	330	2438
0.21873	346	2438
0.19504	360	2438
0.21146	419	2438
0.21050	496	2438
0.21584	334	2439
0.23408	159	2445
0.02011	0	2446
0.25936	18	2446
0.22760	77	2446
0.23673	107	2446
0.21579	139	2446
0.21282	197	2446
0.24251	246	2446
0.21656	382	2446
0.25618	411	2446
0.21716	180	2447
0.49553	435	2452
0.23220	187	2456
0.21783	217	2456
0.24628	27	2457
0.22486	58	2457
0.21355	88	2457
0.21813	208	2457
0.23457	496	2457
0.00882	0	2458
0.20912	19	2458
0.20497	50	2458
0.19674	78	2458
0.19428	109	2458
0.18167	140	2458
0.19349	170	2458
0.17809	199	2458
0.18898	231	2458
0.20700	246	2458
0.20805	261	2458
0.18970	291	2458
0.19509	320	2458

0.17371	334	2458
0.20023	352	2458
0.18486	382	2458
0.21867	411	2458
0.29340	442	2458
0.21022	455	2458
0.24533	18	2460
0.19487	171	2460
0.22939	232	2460
0.03820	0	2463
0.03107	0	2464
0.24993	141	2464
0.27865	56	2467
0.20932	456	2476
0.21343	11	2479
0.22589	41	2479
0.22853	72	2479
0.19230	132	2479
0.23130	146	2479
0.18784	195	2479
0.20341	226	2479
0.19793	255	2479
0.19297	376	2479
0.27289	438	2479
0.22719	447	2479
0.33998	433	2482
0.31203	101	2489
0.35676	437	2489
0.20864	199	2492
0.21429	229	2492
0.37058	443	2492
0.20601	171	2494
0.25031	484	2494
0.42068	101	2498
0.77103	427	2502
0.49086	318	2505
0.20785	332	2506
0.19249	334	2506
0.21992	55	2507
0.30321	433	2507
0.23948	96	2508
0.19610	159	2508
0.20063	223	2513
0.21270	337	2513
0.23678	367	2513
0.26803	430	2513
0.20270	254	2514
0.20200	380	2514
0.32513	442	2514
0.05607	0	2521

0.19800	195	2527
0.19996	120	2530
0.25366	279	2530
0.30791	435	2530
0.49432	279	2547
0.19191	142	2559
0.19736	199	2564
0.21445	357	2567
0.20956	363	2567
0.25210	428	2567
0.20168	193	2574
0.28335	429	2574
0.25317	447	2574
0.21391	193	2575
0.26699	68	2577
0.25018	416	2579
0.40008	444	2579
0.23274	18	2585
0.22989	30	2585
0.19868	84	2585
0.20864	295	2585
0.19944	379	2585
0.19434	382	2585
0.21548	409	2585
0.20596	470	2585
0.60056	440	2586
0.38139	437	2591
0.84544	242	2627
0.64559	101	2655
1.45237	180	2774
0.66767	195	2791
0.90369	325	2821
0.79445	334	2821
0.93292	256	2836
0.79686	375	2836
1.05010	199	2857
0.61794	443	2967
1.17015	178	2978
0.80176	217	2983
0.62210	318	2986
1.48483	456	3002
1.74701	43	3028
0.28443	279	3043
0.26448	416	3053
0.39339	428	3062
0.05757	0	3089
0.20731	159	3109
0.19427	217	3112
0.31597	437	3155
0.20289	142	3166

0.32054	433	3166

Table A.5: Unlearned - Vanilla Saliency - Full Results

SR			GSVT			AFIB			SB		
TF-IDF	Feature	Record	TF-IDF	Feature	Record	TF-IDF	Feature	Record	TF-IDF	Feature	Record
0.55669	0	0	0.00000	0	0	0.01668	0	0	0.00000	0	0
0.00000	1	0	0.78358	92	0	0.00000	1	0	0.42316	25	0
0.69965	352	0	0.93825	300	0	0.52065	139	0	0.55256	84	0
0.61241	1	1	0.10473	332	0	0.01281	303	0	0.38928	151	0
0.33074	259	1	0.41360	410	0	0.44762	327	0	0.47092	273	0
0.59535	2	2	0.36584	421	0	0.01362	106	1	0.36300	311	0
0.44770	260	2	0.09750	164	1	0.31893	125	1	0.34822	496	0
0.54492	3	3	0.46890	255	1	0.24802	143	1	0.37853	28	1
0.40432	46	3	0.56146	374	1	0.01904	409	1	0.49428	63	1
0.41243	267	3	0.41476	207	2	0.24529	105	2	0.42125	151	1
0.57406	5	5	0.33196	424	2	0.01175	212	2	0.25230	403	1
0.61626	58	5	0.08746	160	3	2.00366	384	2	0.34789	488	1
0.50648	295	6	0.47880	240	3	0.01390	15	3	0.28758	1	2
0.49974	84	7	0.54131	325	3	0.36622	91	3	0.40820	34	2
0.42807	8	8	0.64816	359	3	0.28480	148	3	0.31261	172	2
0.51746	53	8	0.10483	495	3	0.31485	165	3	0.27519	425	2
0.48164	96	8	0.12552	6	4	0.24694	284	3	0.32471	111	3
0.56656	259	8	0.08414	26	4	0.01733	318	3	0.26816	332	3
0.36935	188	9	0.09273	38	4	0.27408	481	3	0.22727	365	3
0.38103	410	9	0.00591	324	4	0.42998	20	4	0.34242	203	4
0.14599	10	10	0.11520	331	4	0.01704	121	4	0.44713	238	4
0.30144	18	10	0.07722	359	4	0.47099	211	4	0.31501	321	4
0.31999	61	10	0.08510	371	4	0.40492	306	4	0.38107	12	5
0.31018	105	10	0.09621	380	4	0.58204	481	4	0.27636	109	5
0.32728	134	10	0.06611	206	5	0.45264	82	5	0.30550	352	5
0.33529	197	10	0.07111	441	5	0.50040	274	5	0.27585	411	8
0.32355	293	10	0.20515	497	5	0.01757	30	6	0.20005	38	9
0.30716	314	10	0.17133	15	6	0.20778	67	6	0.22803	69	9
0.29873	470	10	0.13751	50	6	0.22970	228	6	0.24787	91	9
0.27246	11	11	0.15155	65	6	0.01654	333	6	0.18942	105	9
1.20506	186	11	0.11773	156	6	0.01066	136	7	0.21263	191	9
1.23706	320	11	0.04329	152	7	0.37459	340	7	0.32367	317	9
0.60234	12	12	0.44301	237	7	0.40490	428	7	0.23209	345	9
0.46625	445	12	0.32726	298	7	0.01481	242	8	0.32003	380	9
0.63029	362	13	0.36998	318	7	0.59241	422	8	0.24668	452	9
0.27415	17	14	0.25423	381	7	0.76176	433	8	0.28757	488	9
0.29009	62	14	0.10560	484	7	0.27419	83	9	0.26455	34	10
0.28718	111	14	0.15932	499	7	0.00975	348	9	0.29630	70	14
0.31443	155	14	0.18012	66	8	0.90097	374	9	0.24493	108	14
0.29627	160	14	0.21567	135	8	0.77459	413	9	0.34767	143	14
0.29311	200	14	0.08186	316	8	0.01310	454	10	0.21488	178	14
0.30663	246	14	0.31160	358	8	0.65986	498	10	0.26625	335	14

0.31867	377	14	0.26023	8	9	0.44495	223	11	0.22839	429	14
0.28439	424	14	0.06869	148	9	0.01262	60	12	0.29374	331	16
0.27909	445	14	0.49524	274	9	0.36967	230	12	0.37551	88	17
0.27658	474	14	0.10345	480	9	0.33439	313	12	0.26167	145	23
0.07019	15	15	0.20887	42	10	0.01525	363	12	0.24894	275	23
0.33306	19	15	0.28107	335	10	0.01777	166	13	0.20906	365	25
0.29956	47	15	0.22559	407	10	0.00834	469	13	0.34152	59	30
0.26952	87	15	0.33655	431	10	0.01834	272	14	0.23187	296	32
0.30301	139	15	0.24862	471	10	0.28246	364	14	0.25804	333	32
0.32795	240	15	0.07020	144	11	0.01724	378	15	0.19890	204	33
0.41228	16	16	0.23019	159	11	0.25756	68	16	0.19356	55	34
0.36448	173	16	0.40075	480	11	0.23298	131	16	0.55382	192	51
0.35577	224	16	0.33469	497	11	0.21554	143	16	0.47508	331	51
0.32474	284	16	0.29604	174	12	0.00885	181	16	0.50948	405	51
0.39750	490	16	0.26862	193	12	0.29312	275	16	0.44697	429	51
0.11854	17	17	0.06206	308	12	0.34094	287	16	0.53389	335	61
0.25010	46	17	0.08319	140	13	0.21732	344	16	0.47974	427	61
0.25527	88	17	0.24033	156	13	0.01622	287	17	0.38719	438	61
0.25263	125	17	0.20071	216	13	0.85755	46	18	0.44133	239	62
0.23875	134	17	0.16110	235	13	0.39554	190	18	0.62644	457	63
0.23092	179	17	0.17754	284	13	0.30761	248	18	0.41153	490	63
0.24532	223	17	0.21289	494	13	0.34006	335	18	0.61633	433	70
0.29005	262	17	0.18830	126	14	0.31759	413	18	0.72317	75	71
0.24305	270	17	0.14628	140	14	0.33886	97	19	0.21975	45	76
0.24086	477	17	0.17086	214	14	0.01439	302	20	0.32337	365	78
0.38401	18	18	0.09292	304	14	0.26514	350	20	0.80170	37	95
0.48159	252	18	0.11722	147	15	0.01681	105	21	0.89217	73	95
0.48733	294	18	0.15869	161	15	0.01822	408	21	0.64702	233	95
0.47611	382	18	0.10637	171	15	0.41876	104	22			
0.42835	427	18	0.13253	181	15	1.42162	218	22			
0.30425	41	19	0.09106	365	15	1.22222	467	22			
0.31332	221	19	0.13075	300	16	0.37100	177	23			
0.33118	267	19	0.11627	132	17	0.01853	317	23			
0.34904	405	19	0.37634	413	17	0.21792	390	23			
0.35421	452	19	0.11789	296	18	0.25347	25	24			
0.29110	494	19	0.27026	376	18	0.65491	340	25			
0.08241	20	20	0.36586	67	19	0.48556	389	25			
0.33719	115	20	0.07258	128	19	0.01538	332	26			
0.34415	159	20	0.17882	21	21	0.00529	135	27			
0.32768	202	20	0.19233	77	21	0.70067	139	27			
0.34785	248	20	0.01727	124	21	0.01861	438	27			
0.37416	293	20	0.19314	292	21	0.39248	261	28			
0.34060	339	20	0.04479	288	22	0.01874	44	29			
0.31132	386	20	0.04813	120	23	0.28994	64	29			
0.36002	476	20	0.28420	349	23	0.01794	150	30			
0.26088	58	21	0.08082	448	25	0.47475	346	30			
0.28560	77	21	0.30554	491	25	0.01886	256	31			
0.26703	173	21	0.20995	177	26	0.01034	271	34			
0.23670	191	21	0.07932	280	26	0.36628	111	35			
0.27383	280	21	0.24523	292	27	0.29959	420	35			

0.26388	370	21	0.02880	108	29	0.01407	498	39
0.29610	266	22	0.20773	385	33	0.01428	104	41
0.35973	400	22	0.03325	432	33	0.35108	479	42
0.28871	446	22	0.37196	43	34	0.01086	225	45
0.26730	447	23	0.30568	382	34	0.30936	96	47
0.12647	24	24	0.24735	214	35	0.56731	252	48
0.27288	62	24	0.06690	92	37	0.00761	43	49
0.26225	106	24	0.22582	377	37	0.00215	346	49
0.31007	151	24				0.23343	9	50
0.25749	171	24				0.26348	111	50
0.27582	276	24				1.10270	173	52
0.27889	322	24				0.51317	77	55
0.28900	372	24				0.28458	356	58
0.29273	419	24				0.94802	415	58
0.28210	436	24				0.00647	118	64
0.38970	43	25				1.10557	28	70
0.38527	85	25				0.54806	32	74
0.45133	130	25				0.01010	284	77
0.37308	219	25				0.26672	304	77
0.37697	263	25						
0.35258	307	25						
0.39435	355	25						
0.36575	377	25						
0.38976	29	26						
0.40829	153	27						
0.44134	0	28						
0.57531	39	28						
0.62312	354	29						
0.60789	30	30						
0.54914	32	32						
0.26736	247	32						
0.25523	335	32						
0.41845	33	33						
0.42121	149	33						
0.43080	416	33						
0.49556	34	34						
0.24237	198	34						
0.23264	306	34						
0.26069	327	34						
0.23731	448	34						
0.30531	469	36						
0.27007	55	38						
0.75437	91	39						
0.59113	40	40						
0.51762	41	41						
0.35172	151	41						
0.51048	42	42						
0.49397	480	42						
0.51242	3	43						
0.53001	43	43						

0.43572	446	43
0.52935	35	44
0.11155	44	44
0.46586	82	44
0.45199	108	44
0.46105	284	44
0.43958	379	44
0.45643	388	44
0.41810	416	44
0.19005	45	45
0.41673	73	45
0.40048	166	45
0.39679	298	45
0.42589	335	45
0.17267	46	46
0.34662	181	46
0.41562	209	46
0.39924	433	46
0.35487	47	47
0.33390	101	47
0.48380	49	49
0.45306	88	49
0.38003	137	49
0.34827	50	50
0.39322	321	50
0.44704	2	51
0.35941	51	51
0.28169	309	51
0.58628	52	52
0.60012	24	53
0.90200	91	54
0.88259	184	54
0.81153	189	54
0.84036	422	54
0.46642	55	55
0.40439	229	56
0.43522	58	58
0.29356	94	58
0.31659	497	58
0.31867	59	59
0.36228	440	59
0.40984	466	59
0.43594	308	60
0.54655	482	60
0.32988	61	61
0.42177	280	64
0.35894	411	64
0.42835	53	65
0.24848	65	65
0.59155	137	67

0.59736	164	67
0.50288	321	68
0.53383	123	72
0.44513	74	74
0.34101	73	75
0.53977	390	75
0.45945	99	76
0.69302	365	77
0.35571	229	78
0.67446	82	79
0.71368	363	79
0.68044	39	80
0.81940	71	80
0.64720	272	80
0.68662	59	81
0.60970	350	81
0.23105	101	85
0.22851	138	85
0.22374	153	85
0.20617	173	85
0.22150	211	85
0.24925	248	85
0.21934	283	85
0.23371	318	85
0.21332	351	85
0.23943	456	85
0.24579	492	85
0.24506	113	86
0.25395	145	86
0.22831	178	86
0.27246	213	86
0.21691	248	86
0.22234	283	86
0.23979	310	86
0.24788	319	86
0.23044	353	86
0.23493	386	86
0.25723	488	86
0.25084	79	87
0.27034	146	87
0.23471	222	87
0.25801	251	87
0.24766	285	87
0.29483	429	88
0.22910	460	88
0.28143	58	89
0.25091	162	90
0.25983	392	90
0.28546	432	90
0.23939	45	91

0.31177	227	91
0.27752	454	91
0.30086	170	92
0.26476	290	92
0.29668	22	94
0.25304	399	94
0.24885	18	95
0.32071	426	97
0.31518	28	104
0.24685	107	109
0.24491	301	109
0.22427	118	115
0.22048	187	115
0.32671	221	116
0.23279	207	119
0.32316	211	128
0.26515	290	128
0.31043	468	137
0.27180	319	149
0.35789	99	178
0.26437	279	201
0.21520	394	202
0.06534	205	205
0.25713	161	207
0.34448	37	209
0.33855	52	212
0.47087	406	218
0.33962	53	221
0.28639	219	221
0.36918	80	222
0.34419	11	223
0.28413	416	224
0.37903	332	227
0.36566	68	228
0.26307	420	228
0.31912	406	236
0.31644	145	237
0.32188	354	237
0.39105	209	238
0.37943	403	240
0.31384	153	255
0.07581	255	255
0.38505	492	267
0.27773	241	287
0.37207	144	323
1.58716	445	337
1.47459	246	338
0.92327	469	338
0.98747	271	339
0.57025	242	340

0.52270	260	341
0.55934	161	342
0.50757	277	343
0.94678	246	345
0.21134	345	345
0.83279	492	345
0.86473	126	346
0.45982	346	346
0.55249	277	352
0.58053	465	352
0.56061	261	353
0.06112	381	381
0.38443	333	412
0.38666	50	415
0.35093	304	416
0.22626	237	470
0.52815	475	473
0.54598	133	474
0.38642	406	475
0.60341	423	476
0.70653	486	477
0.77358	133	478
0.79505	168	478
0.50255	480	480
0.63781	96	483
0.65405	430	486
0.58594	248	487
1.54121	337	488
0.64572	87	489
0.56536	296	489
0.34956	424	493
0.72889	411	497
0.84819	204	499
0.81839	306	499
0.56620	0	501
0.42484	428	505
0.49338	462	505
0.56898	141	506
0.51362	328	506
0.43198	327	507
0.44357	486	507
0.48132	68	512
0.49892	116	512
0.40868	467	513
0.43541	497	524
0.37697	394	529
0.47352	278	534
0.38317	445	537
0.37401	415	544
0.53807	403	563

0.34378	485	587
0.41203	478	590
0.53486	49	596
0.23802	98	599
0.30160	102	603
0.09979	105	606
0.47017	111	612
1.16288	159	623
1.11483	184	623
0.45420	123	624
0.48973	323	627
0.33780	127	628
0.48563	351	631
0.72113	121	633
0.82548	377	634
0.89208	121	635
0.42142	0	636
0.67221	241	636
0.78400	175	637
0.49835	99	639
1.17627	300	640
1.12611	59	641
1.15010	96	641
0.55601	64	642
0.55154	141	644
0.47401	339	648
0.71655	477	649
0.54720	23	653
0.57420	65	660
0.91238	177	660
0.58229	238	661
0.25407	161	662
0.30562	405	683
0.30191	163	710
0.46042	351	716
0.30887	88	718
0.33833	103	733
0.33070	204	751
0.40446	79	752
0.18388	262	763
0.28922	266	767
0.30820	268	769
0.22388	281	782
0.33479	203	783
0.13553	283	784
0.19665	286	787
0.15821	288	789
0.54036	306	807
0.62203	307	808
0.58066	316	817

0.61933	319	820
0.21164	343	824
0.20748	366	824
0.24299	384	824
0.20952	402	824
0.22648	124	825
0.20362	170	825
0.24700	297	825
0.21615	319	825
0.22608	356	829
0.21726	421	829
0.20963	25	830
0.25294	59	830
0.21526	130	830
0.24252	228	830
0.22393	294	841
0.23650	473	843
0.26569	171	844
0.29374	347	848
0.28708	75	854
0.05742	353	854
0.21867	280	858
0.28181	186	859
0.27695	284	861
0.26828	86	863
0.30000	211	884
0.16276	384	885
0.34370	395	896
0.40240	398	899
0.49826	402	903
0.22107	422	905
0.23577	482	905
0.20551	166	906
0.25133	226	906
0.22940	289	906
0.36526	405	906
0.40799	407	908
0.52094	427	928
0.35799	98	964
0.73699	319	1047
0.45136	475	1055
0.61616	66	1068
0.28629	77	1079
0.21145	61	1080
0.76372	3	1088
0.68215	150	1096
0.59091	120	1107
0.74547	392	1114
1.19030	309	1115
1.09351	420	1115

1.07366	12	1119
0.66286	426	1119
0.93472	270	1173
0.53793	174	1176
1.51324	79	1191
0.42267	213	1215
0.50582	218	1220
0.54750	53	1234
0.36557	189	1237

Appendix B

Long Tables

Model: “model”

```
-----------------------------------------------------------------
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         [(None, 500, 500, 3)]     0
-----------------------------------------------------------------
block1_conv1 (Conv2D)        (None, 500, 500, 64)      1792
-----------------------------------------------------------------
block1_conv2 (Conv2D)        (None, 500, 500, 64)      36928
-----------------------------------------------------------------
block1_pool (MaxPooling2D)   (None, 250, 250, 64)      0
-----------------------------------------------------------------
block2_conv1 (Conv2D)        (None, 250, 250, 128)     73856
-----------------------------------------------------------------
block2_conv2 (Conv2D)        (None, 250, 250, 128)     147584
```

```
-----------------------------------------------------------------

block2_pool (MaxPooling2D)   (None, 125, 125, 128)     0

-----------------------------------------------------------------

block3_conv1 (Conv2D)        (None, 125, 125, 256)     295168

-----------------------------------------------------------------

block3_conv2 (Conv2D)        (None, 125, 125, 256)     590080

-----------------------------------------------------------------

block3_conv3 (Conv2D)        (None, 125, 125, 256)     590080

-----------------------------------------------------------------

block3_pool (MaxPooling2D)   (None, 62, 62, 256)       0

-----------------------------------------------------------------

block4_conv1 (Conv2D)        (None, 62, 62, 512)       1180160

-----------------------------------------------------------------

block4_conv2 (Conv2D)        (None, 62, 62, 512)       2359808

-----------------------------------------------------------------

block4_conv3 (Conv2D)        (None, 62, 62, 512)       2359808

-----------------------------------------------------------------

block4_pool (MaxPooling2D)   (None, 31, 31, 512)       0

-----------------------------------------------------------------

block5_conv1 (Conv2D)        (None, 31, 31, 512)       2359808

-----------------------------------------------------------------

block5_conv2 (Conv2D)        (None, 31, 31, 512)       2359808

-----------------------------------------------------------------

block5_conv3 (Conv2D)        (None, 31, 31, 512)       2359808
```

```
_________________________________________________________________

block5_pool (MaxPooling2D)   (None, 15, 15, 512)       0

_________________________________________________________________

flatten (Flatten)            (None, 115200)            0

_________________________________________________________________

dense (Dense)                (None, 1024)              117965824

_________________________________________________________________

dense_1 (Dense)              (None, 512)               524800

_________________________________________________________________

dense_2 (Dense)              (None, 4)                 2052

=================================================================

Total params: 133,207,364

Trainable params: 118,492,676

Non-trainable params: 14,714,688

_________________________________________________________________
```